(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－120206

(43)公開日　平成11年(1999)４月30日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３１０Ｄ	
17/27		15/38	Ｄ	

審査請求　未請求　請求項の数１　書面　外国語出願　（全 43 頁）

(21)出願番号　特願平10－223557

(22)出願日　平成10年(1998)７月２日

(31)優先権主張番号　０５１５５８

(32)優先日　1997年７月２日

(33)優先権主張国　米国（ＵＳ）

(31)優先権主張番号　１００１８９

(32)優先日　1998年６月18日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　590000798
ゼロックス　コーポレイション
ＸＥＲＯＸ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国　06904－1600　コネティカット州・スタンフォード・ロング　リッチ　ロード・800

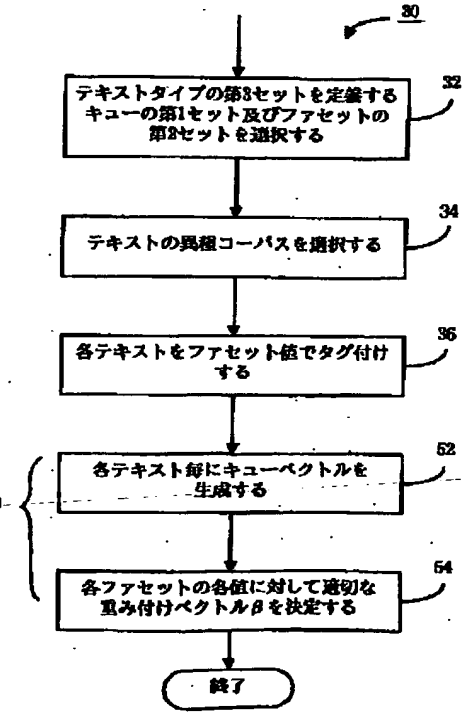
(72)発明者　ジョフリー　ディー．ナンバーグ
アメリカ合衆国　94114　カリフォルニア州　サンフランシスコ　グランド　ビュー　アベニュー　678

(74)代理人　弁理士　中島　淳　（外１名）

最終頁に続く

(54)【発明の名称】　タグ付けされていないテキストの外観特徴を使用したテキストジャンルの自動決定方法及び装置

(57)【要約】
【課題】　マシン可読形式のタグ付されていないテキストのジャンルをプロセッサを用いて識別する方法を提供する。
【解決手段】　この方法は、容易に計算可能である非構造的な表面キューの第１セットがテキストにおいて発生する回数を表すキューベクトルをテキストから生成することによって開始する。この後プロセッサは、キューベクトルと、第１のテキストジャンルに関連する重み付けベクトルとを用いて、このテキストが第１のテキストジャンルのインスタンスであるか否かを決定する。



1

【特許請求の範囲】
【請求項１】　テキストの構造分析を行わずに、マシン可読形式のタグ付けされていないテキストのテキストジャンルをプロセッサを用いて識別する方法であって、
ａ）非構造的な表面キューの第１セットが前記テキストにおいて発生する回数を表すキューベクトルを前記テキストから生成するステップと、
ｂ）前記キューベクトルと、第１のテキストジャンルに関連する重み付けベクトルを用いて、前記テキストが前記第１のテキストジャンルのインスタンスであるか否かを決定するステップと、
を含む、テキストジャンル識別方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は計算言語学に関する。
【０００２】
【従来の技術及び発明が解決しようとする課題】「ジャンル」という言葉は通常、「テキストの種類」の代わりに用いる文学的な言葉として機能する。テキストジャンルは、テキストトピック（題目）及び文書ジャンルの関連概念とは異なる。テキストジャンル及びテキストトピックは、互いから完全に独立してはいない。新聞に記載の話、小説及び科学的な記事などの顕著なテキストジャンルは主に、異なる範囲のトピックを扱っている。しかし、これらのテキストジャンルの各々におけるトピックの共通性は非常に広く抽象的である。更に、単一のトピックに関連する大量のテキストの集まりはどれも１つより多くのテキストジャンルの作品を殆ど常に含み、よってこれらの間の形式的な類似点は語彙アイテムの存在に限られる。概念としてのテキストジャンルは文書ジャンルとは無関係であるが、これら２つのジャンルのタイプは濃密な機能的相互依存と歴史的に密接に関連して発達している。例えば、単一のテキストジャンルはいくつかの文書ジャンルと関連しうる。ショートストーリーを雑誌又は選集に掲載したり、又は小説を複数部分に分けて連続出版したり、小説をハードカバー、そして後にペーパーバックとして再出版したりすることができる。同様に、新聞のような文書ジャンルは、特集記事、コラム、失恋した人へのアドバイス及びクロスワードパズルなど、いくつかのテキストジャンルを含むことができる。これらのテキストジャンルは、「昨日」及び「ローカル」のような文脈に依存する単語の使用を許容する新聞に現れなければ、現在のように読まれていない可能性がある。これらが密接して関連しているために、文書ジャンルの物質的な特徴がテキストジャンルを示すことが多い。例えば、新聞はあるフォントを「ハードニュース（政治・経済・国際関係などに関するニュース）」の見出しに使用して別のフォントを分析の見出しに使用したり、定期刊行物は用紙（ｐａｐｅｒｓｔｏｃｋ）によってそのトピック内容を示したり、ビジネスレター及び私信をページのレイアウトに基づいて区別したりすることができる。異種のデジタルテキストの集まりから関連テキストを検索するのが難しいことが多いのは、デジタル化によってテキスト及び文書ジャンルに関連するこれらの物理的な手掛かりが取り除かれてしまうためである。

2

【０００３】公と私、ジェネラリストとスペシャリスト、仕事と休養などのテキストジャンル間の境界は、社会生活が別個の役割及び行動に分かれていることを反映する。ジャンルは、文書を解釈可能にする情況を提供するため、ジャンルは内容に劣らずユーザの関連概念を形成する。例えば、スーパーコライダー（超衝突装置：ｓｕｐｅｒｃｏｌｌｉｄｅｒ）又はナポレオンに関する情報を求めている研究者は、内容と同じくらいテキストジャンルに注意する。研究者は、出所の内容だけでなく、その出所が学術雑誌に記載されているか又は一般雑誌に記載されているかということも知りたいと思うであろう。
【０００４】最近まで、情報検索及びテキスト分類の研究は、テキストジャンルではなくトピックの識別に殆ど独占的に焦点を当ててきた。テキストジャンルの識別が殆ど研究されなかった理由は２つある。第１に、従来のプリントベース文書の世界ではジャンル分類の必要性がみられなかった。何故なら、この世界では、ジャンルは本質的に、又は画一的な文脈上の特徴によって明確に示されているからである。低温融合に関する記事を探しに図書館を訪ねた科学者は、どうやって定期刊行雑誌の記事に研究を制限するかを心配しなくてもよい。何故なら、定期刊行雑誌は一般的な科学雑誌と区別できるように目録が作られ、書架に置かれているからである。第２に、オンラインのテキストデータベースを用いた迅速な情報検索作業は、百科事典又は新聞のデータベースのように、テキストジャンルが外的に統制される小さく比較的同種のデータベースに焦点を当てていた。テキストジャンル間の境界が示されていないことが多い大きな異種のテキストデータベースによって、テキストのジャンル分類の重要性が強調される。トピックベースの検索ツールのみでは、大きな異種データベースを検索する際に読者の興味の対象物の範囲を適切に選択することができない。
【０００５】ジャンル分類のアプリケーション（用途）は、情報検索の分野に限られていない。いくつかの言語学技術も、このアプリケーションから利益を得ることができる。単語の意味の分布はジャンルによって大幅に異なることが既知であるため、自動的な文（センテンス）の部分のタガー（タグを付けるもの）及び意味のタガーは共にジャンル分類から利益を得ることができる。
【０００６】書籍の分類の論述はアリストテレスまでさかのぼる。ジャンルに関する文献は分類スキーム及びシステムが豊富であり、そのうちのいくつかを単純な属性

システムとして分析することができる。これらの論述はあいまいで、牧歌又は小説のような文学的形式や、これよりも程度が小さいが新聞の犯罪レポート又はラブレターのようなパラ文学的な（ｐａｒａｌｉｔｅｒａｒｙ）形式に専ら焦点を当てる傾向にある。分類の論述は、年次報告、Ｅメール通信及び科学的なアブストラクトなど、文学的ではないテキストのタイプを無視する傾向にある。更に、これらの論述のうちで、ジャンルを区別するアブストラクトディメンション（摘要の範囲）をテキストのあらゆる形式的特徴に関連づける努力をしているものは１つもない。

【０００７】テキストのジャンル分類の量化方法に具体的に関連する唯一の言語学研究は、バイバー（Ｄｏｕｇｌａｓ　Ｂｉｂｅｒ）の研究である。彼の研究は、以下を含む：″Ｓｐｏｋｅｎ　ａｎｄ　Ｗｒｉｔｔｅｎ　Ｔｅｘｔｕａｌ　Ｄｉｍｅｎｓｉｏｎｓ　ｉｎ　Ｅｎｇｌｉｓｈ：Ｒｅｓｏｌｖｉｎｇ　ｔｈｅ　Ｃｏｎｔｒａｄｉｃｔｏｒｙ　Ｆｉｎｄｉｎｇｓ″（Ｌａｎｇｕａｇｅ、６２（２）：３８４－４１３，１９８６）；″Ｖａｒｉａｔｉｏｎ　Ａｃｒｏｓｓ　Ｓｐｅｅｃｈａｎｄ　Ｗｒｉｔｉｎｇ″（Ｃａｍｂｒｉｄｇｅ　Ｕｎｉｖｅｒｓｉｔｙ　Ｐｒｅｓｓ、１９８８）；″Ｔｈｅ　Ｍｕｌｔｉｄｉｍｅｎｓｉｏｎａｌ　Ａｐｐｒｏａｃｈ　ｔｏ　Ｌｉｎｇｕｉｓｔｉｃ　Ａｎａｌｙｓｅｓ　ｏｆ　Ｇｅｎｒｅ　Ｖａｒｉａｔｉｏｎ：Ａｎ　Ｏｖｅｒｖｉｅｗ　ｏｆ　Ｍｅｔｈｏｄｏｌｏｇｙ　ａｎｄ　Ｆｉｎｄｉｎｇ″（Ｃｏｍｐｕｔｅｒｓ　ｉｎ　ｔｈｅ　Ｈｕｍａｎｉｔｉｅｓ、２６（５－６）：３３１－３４７、１９９２）；″Ｕｓｉｎｇ　Ｒｅｇｉｓｔｅｒ－Ｄｉｖｅｒｓｉｆｉｅｄ　Ｃｏｒｐｏｒａ　ｆｏｒ　Ｇｅｎｅｒａｌ　Ｌａｎｇｕａｇｅ　Ｓｔｕｄｉｅｓ″（Ｕｓｉｎｇ　ＬａｒｇｅＣｏｒｐｏｒａ、第１７９－２０２頁（Ｓｕｓａｎ　Ａｒｍｓｔｒｏｎｇ　編集）（１９９４））；及びフィネガンＥｄｗａｒｄ　Ｆｉｎｅｇａｎ）と共著の″Ｄｒｉｆｔ　ａｎｄ　ｔｈｅ　Ｅｖｏｌｕｔｉｏｎ　ｏｆ　ＥｎｇｌｉｓｈＳｔｙｌｅ：Ａ　Ｈｉｓｔｏｒｙ　ｏｆ　Ｔｈｒｅｅ　Ｇｅｎｒｅｓ″（Ｌａｎｇｕａｇｅ、６５（１）：９３－１２４、１９８９）。バイバーの研究は記述的であり、各々が利用する傾向にある言語学的特徴のタイプに従ってテキストジャンルを機能的に区別することを目的としている。バイバーは、「学究散文」及び「一般的なフィクション」など、手作業によって多数の別個のジャンルに分けられたコーパスから始めている。次に、通常は３つか５つである、テキストのいくつかの「ディメンション」又は要素に沿ってこれらのジャンルをランク付けする。バイバーは、殆どが統語的又は語彙的なものである言語学的特徴のセットにこの要素分析を適用することによって要素を個性化している。これらの要素には、例えば過去時制の動詞、過去分詞節及び″ｗｈ－″から始まる質問などが含まれる。次にバイバーは、言語学者が各要素の個々の構成部分に割り当てて用いてきた談話機能（例えば、「情報を与えるｖｓ関係のある」ディメンション、「説話的ｖｓ非説話的」ディメンションなどとして）を抜粋することにより、一般的な意味又は機能を要素に割り当てている。ジャンルに従って個々のテキストを分類する際に、これらの要素はその有用性に従って個性化されるのではないことに注意する。所与の要素又は要素のセットに対してあらゆるテキストが受け取るスコアは、そのジャンルほど多くの情報を与えるものではない場合がある。何故なら、あらゆる個々の要素に関連するジャンル間に大幅な重複があるからである。

【０００８】カールグレン（Ｊｕｓｓｉ　Ｋａｒｌｇｒｅｎ）及びカッテイング（Ｄｏｕｇｌａｓｓ　Ｃｕｔｔｉｎｇ）は、″Ｒｅｃｏｇｎｉｚｉｎｇ　Ｔｅｘｔ　Ｇｅｎｒｅｓ　ｗｉｔｈ　Ｓｉｍｐｌｅ　Ｍｅｔｒｉｃ　Ｕｓｉｎｇ　Ｄｉｓｃｒｉｍｉｎａｎｔ　Ａｎａｌｙｓｉｓ″（Ｐｒｏｃｅｅｄｉｎｇｓ　ｏｆ　Ｃｏｌｉｎｇ’９４、第ＩＩ巻、第１０７１－１０７５頁、１９９４年８月）において、バイバーの結果の一部をジャンルの自動分類に適用するための努力を述べている。彼らもまた、手作業で分類したテキストのコーパス、即ちブラウンコーパスから始めている。ブラウンコーパスをまとめた人々はこの分類を総称的なものと述べているが、教養のある読者が認識するテキストとジャンルとの間の適合はおおよそにすぎない。カールグレン及びカッティングは、語彙特徴又は分布特徴のいずれかを用いる。語彙特徴は第１人称代名詞の総数及び現在時制の動詞の総数を含み、分布特徴は長い単語の総数及び単語当たりの平均文字数を含む。彼らは、句読レベル又は文字レベルの特徴を使用しない。この２人の著者は、判別分析を用いてテキストを様々な数のカテゴリーに分類する。カールグレン及びカッティングが手作業で割り当てたカテゴリーの数に等しい数の機能を用いたとき、自動的に得たカテゴリーと手作業で分類したカテゴリーとの間の適合は５１．６％であった。機能の数を減少させ、コーパスのカテゴリーを再構成することによって、彼らは実施を改良した。カールグレン及びカッティングは、このような方法が情報検索の目的に有用であるか定かではないと考えており、以下のように述べている：「自動的に得たカテゴリーを使用する際の問題は、たとえこれらのカテゴリーがデータによって支持されているという意味で実質的なものであっても、この技術を検索ツールにおいて使用することが目的である場合、これらのカテゴリーは熱心でない素人に対して説明することが難しくなりうる、ということである。」更に、ブラウンコーパスの特有の「ジャンル」が、ユーザが情報検索のタスクに関連して見出すカテゴリーとどの程度一致するかが明らかではない。

【０００９】ナンバーグ（Ｇｅｏｆｆｒｅｙ　Ｎｕｎｂ

ｅｒｇ）及びヴィオリ（Ｐａｔｒｉｚｉａ　Ｖｉｏｌｉ）は、″Ｔｅｘｔ，Ｆｏｒｍ　ａｎｄ　Ｇｅｎｒｅ″（Ｐｒｏｃｅｅｄｉｎｇｓ　ｏｆ　ＯＥＤ’９２、第１１８－１２２頁、１９９２年１０月）において、ジャンルの認識が、情報検索のタスク及び自然言語処理のタスクに重要であることを示唆している。これらの著者は、テキストのジャンルをクラスではなく属性として処理することができると提案している。しかし、彼らは識別を達成できる態様に関する具体的な提案を提供していない。

【００１０】

【課題を解決するための手段】マシン可読でタグ付けされていないテキストのジャンルを自動的に識別する本発明の方法は、様々な利点を提供する。簡潔に説明すると、プロセッサによって実施される本方法は、テキストからキューベクトルを生成することによって始まる。キューベクトルは、容易に計算可能である非構造的な表面キューの第１セットがテキストにおいて発生する回数を表す。その後、プロセッサは、キューベクトルと、第１のテキストジャンルに関連する重み付けベクトルとを用いて、テキストが第１のテキストジャンルのインスタンスであるか否かを決定する。

【００１１】

【発明の実施の形態】図１は、命令１００を実行することによって本発明の方法が行われるコンピュータシステム１００をブロック図で示している。本発明の方法はコンピュータシステム１０の動作を変え、マシン可読形式でシステムに提供されるタグ付けされていないテキストのテキストジャンルを自動的に決定することができるようにする。命令１００によって、テキストの構造分析、単語のステミング（語幹化：ｗｏｒｄ　ｓｔｅｍｍｉｎｇ）又は品詞のタグ付けを行わずにテキストジャンルの分類を行うことができる。命令１００は、構造ベースの特徴よりもより迅速に計算することができる新しい表面レベルのキュー又は特徴に依存する。簡潔に述べると、命令１００に従って、コンピュータシステム１０はテキストを分析し、このテキスト内の各表面キューの発生回数を決定してキューベクトルを生成する。次にコンピュータシステム１０は、テキストが特定のテキストジャンル及び／又はファセットのインスタンスであるか否かを、キューベクトルと、特定のテキストジャンル及び／又はファセットに関連する重み付けベクトルとを用いて決定する。命令１００は、図４に関連して詳しく説明される。コンピュータシステム１０は、学習（トレーニング）命令５０を用いて各テキストジャンル及び／又はファセットに適切な重み付けベクトルを決定する。これは、図３に関連して詳しく説明される。

【００１２】Ａ．テキストジャンルを自動的に決定するコンピュータシステム

命令５０及び１００をより詳しく説明する前に、これらの命令を実行するコンピュータシステム１０について説明する。図１に示されるように、コンピュータシステム１０は情報をコンピュータユーザに視覚的に表示するモニタ１２を含む。また、コンピュータシステム１０はプリンタ１３を介してコンピュータユーザに情報を出力する。コンピュータシステム１０は、データを入力する複数の経路をコンピュータユーザに提供する。キーボード１４を打つことによって、コンピュータユーザはコンピュータシステム１０に入力データを入力することができる。マウス１６を動かすことによって、コンピュータユーザはモニタ１２に表示されたポインタを動かすことができる。また、コンピュータユーザは、スタイラス２０又はペンで電子タブレット１８に書き込むことによってコンピュータシステム１０に情報を入力することもできる。あるいは、フロッピーディスクなどの磁気媒体をフロッピーディスクドライブ２２に挿入することにより、コンピュータユーザは磁気媒体に記憶されたデータを入力することができる。スキャナー２４によって、コンピュータユーザはハードコピー文書のマシン可読バージョン、例えばＡＳＣＩＩを生成することができる。

【００１３】プロセッサ１１は、コンピュータシステム１０の動作の制御及び統制を行い、コンピュータユーザのコマンドを実行する。プロセッサ１１は、メモリ２８又はディスクドライブ内のフロッピーディスクに電子的に記憶された命令５０及び１００などの命令を実行することにより、各ユーザのコマンドに応答する適切な動作を判断し、これを行う。通常、プロセッサ１１のための動作命令は固体メモリに記憶され、これによって命令に頻繁かつ迅速にアクセスすることができる。メモリの具現に使用することができる半導体論理デバイスには、読出し専用メモリ（ＲＯＭ）、ランダムアクセスメモリ（ＲＡＭ）、ダイナミックＲＡＭ（ＤＲＡＭ）、プログラマブルＲＯＭ（ＰＲＯＭ）、消去可能型ＰＲＯＭ（ＥＰＲＯＭ）及びフラッシュメモリなどの電気的書き込み可能型ＲＯＭ（ＥＥＰＲＯＭ）が含まれる。

【００１４】Ｂ．テキストのジャンル、ファセット及びキュー

コンピュータシステム１０は命令５０及び１００に従って、構造分析、ステミング、解析又は意味もしくは品詞のタグ付けをまだ行っていないトークン化されたマシン可読テキストのテキストジャンルを決定する。本明細書中に使用されるように、「テキストジャンル」とは、テキストが示す直接のトピックによって直接に生じたものではないいくつかの形式キュー又は共通属性に機能が関係していることを条件として、いくつかの共通の通信目的特徴又は他の機能的特徴（ｔｒａｉｔ）によって定義されるテキストの広く認識された任意のクラス（種類）をいう。テキストのクラスが広く認識されていることにより、一般の人々は解釈原理の特徴的なセットを用いてクラスのテキストを解釈することができる。本明細書中

に使用されるように、テキストジャンルは文（センテンス）のジャンルのみに適用する。即ち、テキストジャンルは、句読及びパラグラフなどのテキストカテゴリーインジケータの十分なレパートリーを利用するストリングのような文（単数及び複数）を主に介して伝わるジャンルのみに適用する。従って、本発明では、航空路のスケジュール、株式の表及びコマ漫画はテキストジャンルとして認識されない。また、本発明は会話のジャンルもテキストジャンルとして認識しない。テキストジャンルによって定義されるクラスは拡張可能であることが好ましい。従って、本発明では、ジェーン・オースティン（Ｊａｎｅ　Ａｕｓｔｅｎ）によって書かれた小説のクラスは拡張可能ではないため、好適なテキストジャンルではない。

【００１５】命令５０及び１００の方法は、テキストジャンルをファセットの集まりとみなす。各ファセットは、キュー又は特徴と呼ばれる計算可能な言語学特性の特徴的なセットと関連しており、これらはテキストの形式の表面レベル特徴から観察することができる。これらのキューを使用して、各ファセットは一定の実用的な対＊

＊象物に応えるテキストのクラスを区別する。１つのファセットが複数のジャンルに関連する場合があるため、ファセットはテキストジャンルを間接的に識別する傾向にある。どのテキストジャンルもファセットの特定のクラスタとして定義することができるため、本発明の方法は、他のアプローチと同じ正確さであるが以前にはなかった新規のテキストジャンルを容易に追加することができるという利点を有してテキストジャンル及びスーパージャンルを識別することができる。

【００１６】ファセットの概念を更に定義しようとする代わりに、例示的な具体例をいくつか説明する。読者（ａｕｄｉｅｎｃｅ）ファセットは、広範囲のテキストと、より限られた読者にむけられたテキストとを区別する。長さファセットは、短いテキストと長いテキストの区別をする。組織又は匿名及び個人によって書かれたテキストの区別は、著者ファセットによって表される。下記のリストは、これらの値が明確でないときの他のファセット及びその値である。ファセットは２値でなくてもよいことに注意する。

ファセット名	可能な値
１．日付	あり／なし
２．説話的	Ｙｅｓ／Ｎｏ
３．説得的（議論的）／記述的（教育的）	
４．フィクション／ノンフィクション	
５．法的	Ｙｅｓ／Ｎｏ
６．科学及び技術的	Ｙｅｓ／Ｎｏ
７．知的水準　平俗	Ｙｅｓ／Ｎｏ
（Ｂｒｏｗ）中	Ｙｅｓ／Ｎｏ
高	Ｙｅｓ／Ｎｏ

【００１７】他のファセットを定義して、本発明と矛盾せず上記リストのファセットに追加することができる。テキストジャンルを定義するのに全てのファセットを用いる必要はなく、テキストジャンルを単一のファセットで定義することができる。下記のリストは、前述のファセット及び値を用いて定義することができる、従来認識されているテキストジャンルのいくつかの例にすぎない。

１．　新聞の報道

ａ．読者	広範囲
ｂ．日付	あり
ｃ．説得的	記述的
ｄ．説話的	Ｙｅｓ
ｅ．フィクション	Ｎｏ
ｆ．知的水準	平俗
ｇ．著者	記名なし
ｈ．法的	Ｎｏ

２．　論説の意見

ａ．読者	広範囲
ｂ．日付	あり
※ｃ．説得的	Ｙｅｓ
ｄ．説話的	Ｙｅｓ
ｅ．フィクション	Ｎｏ
ｆ．知的水準	平俗
ｇ．著者	記名あり
ｈ．科学及び技術的	Ｎｏ
ｉ．法的	Ｎｏ

３．　市場分析

ａ．読者	広範囲
ｂ．日付	あり
ｃ．説得的	記述的
ｄ．説話的	Ｎｏ
ｅ．フィクション	Ｎｏ
ｆ．知的水準	高
ｇ．著者	組織
ｈ．科学及び技術的	Ｙｅｓ
ｉ．法的	Ｎｏ

４．　Ｅメール

ａ．読者	受取人
ｂ．日付	あり

9

ｃ．フィクション　　Ｎｏ
ｄ．知的水準　　平俗
ｅ．著者　　記名あり

【００１８】テキストジャンルがファセットのグループに分解するように、ファセットも本方法に従った表面レベルのキューに分解する。本発明の表面レベルキューは、単語のステミング、解析、又は意味もしくは品詞のタグ付けなどの構造分析を全く行わずにトークン化されたＡＳＣＩＩテキストを用いて計算することができるため、本発明の表面レベルキューは従来の特徴とは異なる。本発明に関連するのは、大抵はテキスト内のこれらの表面レベルキューの発生回数（頻度）である。表面レベル又は形式キューのいくつかのタイプを下記に定義できるが、これらに限定されない：数／統計、句読、構造、式文、語彙及び逸脱。方式タイプのキューは、従来特定のテキストジャンルに関連するコロケーション又は定着した表現である。例えば、おとぎ話は″Ｏｎｃｅ　ｕｐｏｎ　ａ　ｔｉｍｅ（むかしむかし）″で始まり、聖母マリアの讃歌は「ヘイルメアリー（Ｈａｉｌ　Ｍａｒｙ　：聖母マリアに捧げる祈り）」で始まる。他の式文は、法律文書、認可承諾書などを示す。語彙タイプのキューは、テキストジャンルを示すことができる一定の語彙アイテムの回数に関連する。例えば、Ｍｒ．、Ｍｒｓ．及びＭｓ．などの習慣的な敬称用語がニューヨークタイムズの記事に使用されており、「昨日」及び「ローカル」などの単語が新聞の報道に頻繁に使用されている。更に、″ｉｔ’ｓ　ｐｒｅｔｔｙ　ｍｕｃｈ　ａ　ｓｎａｐ″などのフレーズを使用する場合、テキストが例えば百科事典の記事の一部ではないことを示している。いくつかの語彙アイテムの使用は、いくつかのテキストジャンルのトピック及び修辞学的な共通属性によって保証される。構造的な特徴は従来技術において既知であるが、その殆どの計算にはタグ付けされたか又は十分に解析されたテキストが必要である。ストリング認識が可能であるこれら２つの新しい表面レベル構造キューは、本発明によって定義される。句読タイプのキューは、テキスト内の句読的特徴の総数である。このタイプのキューは以前に使用されていないが、これらは有意であり、非常に多いため、テキストジャンルの有用なインジケータとして機能することができる。例えば、クエスチョンマークの総数が多ければ、テキストは読者を説得しようとしていることを示す可能性が高い。特定のテキスト内の表面レベル特徴の回数を測定する殆どの他のキュータイプとは対照的に、逸脱タイプのキューは単位サイズ内の逸脱に関連する。例えば、逸脱キューを使用して、テキストジャンルによって変化しうる特徴である文及びパラグラフの長さの変化を追跡することができる。キューのタイプは、テキストの特徴を示すために測定することができる表面レベルの特徴の種類を示唆するために説明したにすぎず、キューのタイプの特徴付けは本発明にとって重要ではない。定義することができるキューの数は、理論的に無制限である。使用可能なキューのほんのいくつかを例示的な目的で下記に列挙する。

10

Ａ．句読のキュー
１．ログ（コンマの総数（カウント）＋１）
２．平均値（コンマ／文）／記事
３．平均値（ダッシュ／文）／記事
４．ログ（クエスチョンマークの総数＋１）
５．平均値（クエスチョンマーク／文）／記事
６．ログ（ダッシュの総数＋１）
７．ログ（セミコロンの総数＋１）
Ｂ．ストリング認識が可能な構造のキュー
１．″ａｎｄ″、″ｂｕｔ″及び″ｓｏ″で始まる文／記事
２．副詞＋コンマで始まる文／記事
Ｃ．式文のキュー
１．″Ｏｎｃｅ　ｕｐｏｎ　ａ　ｔｉｍｅ．．．″
Ｄ．語彙のキュー（他の指示がない限りトークンの総数のみを示す）
１．″Ｍｒ．、Ｍｒｓ．″などの略称
２．頭文字語
３．法助動詞
４．動詞″ｂｅ″の形式
５．暦ー曜日、月
６、７．大文字ー大文字で始まる文ではない初めの単語のタイプ及びトークン数
８．文字数
９、１０．短縮タイプ及びトークン数
１１、１２．″ｅｄ″で終わる単語のタイプ及びトークン数
１３．数式
１４．動詞″ｈａｖｅ″の形式
１５、１６．ハイフン付きの単語のタイプ及びトークン数
１７、１８．多音節語のタイプ及びトークン数
１９．単語″ｉｔ″
２０、２１．ラテン語の接頭辞及び接尾辞のタイプ及びトークン数
２２、２３．６文字よりも多い単語のタイプ及びトークン数
２４、２５．１０文字よりも多い単語のタイプ及びトークン数
２６、２７．３つより多い単語句（Ｔｈｒｅｅ＋ｗｏｒｄ　ｐｈｒａｓｅｓ）のタイプ及びトークン数
２８、２９．″１ｙ″で終わる多節語のタイプ及びトークン数
３０．明白な否定語
３１、３２．少なくとも１つの数字を含む単語のタイプ及びトークン数
３３．左かっこ

３４、３５．前置詞のタイプ及びトークン数
３６．第１人称単数の代名詞
３７．第１人称複数の代名詞
３８．引用符の対
３９．ローマ数字
４０．″ｔｈａｔ″のインスタンス
４１．″ｗｈｉｃｈ″のインスタンス
４２．第２人称複数の代名詞
Ｆ．逸脱のキュー
１．文の標準の長さからの逸脱（単語数）
２．単語の標準の長さからの逸脱（文字数）
３．句読点間のテキストセグメントの標準の長さからの逸脱（単語数）
４．平均値（文字／単語）／記事

【００１９】約４００のテキストのコーパスを用いた事前試行の結果として、図２の表１はいくつかの表面レベルのキューがファセット／テキストジャンルによって変化しうる態様を示している。（この試行は、上記のようにテキストジャンルを分解せず、いくつかのテキストジャンルを単一のファセットとみなした。双方のアプローチは本発明と矛盾しない。前述のように、テキストジャンルを単一のファセットによって定義することができる。）例えば、このコーパス内で、新聞の報道は１つの記事当たり１．２個のセミコロンしか含まなかったが、法律文書は４．７８個含んだ。同様に、テキスト当たりのダッシュの数は、新聞の報道、論説の意見及びフィクションにおいて異なっていた。

【００２０】異なるキュー値にどの位の重みを付けるべきか？換言すると、特定のファセット又はテキストジャンルのキュー値又はキュー値のセットはどれだけ密接に相関しているのか？人間が判断する事柄であるテキストジャンルのファセット値への分解とは対照的に、この質問に対する答えは人間が判断する事柄ではない。ファセットに従って各キューに合った重みを決定するには、図３に関連して後述する学習が必要である。

【００２１】Ｃ．キューの重みを決定するための学習
図３は、各キュー毎にキューの重みを決定するための学習方法３０をフロー図で示している。学習方法３０は完全に自動ではなく、ステップ３２、３４及び３６はマニュアルで実行され、命令５０のステップはプロセッサによって実行される。命令５０は、固体メモリ又はフロッピーディスクドライブ内に配置したフロッピーディスクに記憶させることができ、ＬＩＳＰ及びＣ＋＋を含むあらゆるコンピュータ言語で実現させることができる。

【００２２】学習方法３０は１セットのキュー及び別の１セットのファセットの選択で始まり、これらを使用して広く認識された１セットのテキストジャンルを定義することができる。ステップ３２において約５０～５５個の表面レベルキューを選択することが好ましいが、これよりも少ないか又は多い数を本発明と矛盾せず使用することができる。また、語彙及び句読タイプの表面レベルキューの数を選択することが好ましい。ユーザは定義される各ファセットに表面レベルキューを全て組み込むことができるが、これは必須ではない。ステップ３２において任意の数のファセットを定義し選択できるが、ユーザは何らかの数のファセットを定義しなければならない。反対に、後述するように、ファセットそのものが多数のアプリケーションにおいて有用であるため、ユーザはこの時点ではテキストジャンルを定義しなくてよい。

この後、ステップ３４においてユーザはテキストの異種コーパスを選択する。テキストジャンルが定義されていない場合、選択されるコーパスは、選択されるテキストジャンル又はファセットの各々において約２０個のインスタンスを含むことが好ましい。通常はＡＳＣＩＩであるデジタル又はマシン可読形式でない場合、命令５０に進む前にコーパスを変換してトークン化しなければならない。ファセット、表面レベルキュー及び異種コーパスの選択後、ユーザはステップ３６においてマシン可読ファセット値をコーパスのテキストの各々に関連づける。この後に、ユーザは残りの学習タスクをコンピュータシステム１０に引き継ぐ。

【００２３】命令５０はステップ５２から始まる。このステップにおいて、プロセッサ１１はコーパスの各テキスト毎にキューベクトルＸを生成する。キューベクトルは、選択されたキューの各々に対して１つの値を有する多次元のベクトルである。プロセッサ１１は、特定のテキスト内にみられる関連した表面レベルの特徴に基づいて、各キューの値を決定する。選択されたキューの定義に基づいてキュー値を決定する方法は当業者には明らかであるため、本明細書では詳しく説明しないことにする。これらの方法にはテキストの構造分析又はタグ付けが必要ではないため、プロセッサ１１はステップ５２においてキュー値を決定するために比較的わずかな計算を行うだけでよい。

【００２４】ステップ５４において、プロセッサ１１はファセット値に従って各キューに付けられるべき重みを決定する。即ち、ステップ５４において、プロセッサ１１は各ファセットに対して重み付けベクトルβを生成する。キューベクトルＸのように、重み付けベクトルβは選択されたキューの各々に対して１つの値を有する多次元ベクトルである。ロジスティック回帰を含む多数の数学的アプローチを使用して、コーパスのキューベクトルから重み付けベクトルを生成することができる。ロジスティック回帰を用いて、プロセッサ１１はステップ５２で生成されたキューベクトルを同一のキューベクトルのセットに分割する。次に、各２値ファセットに対して、プロセッサ１１は同一キューベクトルの各セットに対するログ奇関数を解く。ログ奇関数ｇ（ψ）は、下記のように表される。

$$g(\psi) = \log(\psi / 1-\psi) = X\beta$$

式中、ψはファセット値が真であるベクトルの割合であり、$1-\psi$はファセット値が偽であるセット内のベクトルの割合である。

【００２５】ファセット値の先のタグ付けは、同一のキューベクトルを有するテキストの各セット内に各ファセット値を有するテキストの数を示すため、プロセッサ１１はψ及び$1-\psi$の値を決定することができる。従って、プロセッサ１１は、同一キューベクトルのセット、既知のψ値のセット、$1-\psi$値のセット及びキューベクトル値のセット全てによって定義される連立方程式のシステムを解くことにより、各２値ファセットのための重み付けベクトルβの値を決定することができる。ロジスティック回帰は公知であり、本明細書では詳しく説明しないことにする。ロジスティック回帰のより詳細な論述に関しては、本明細書に援用されるマッカラー（ＭｃＣｕｌｌａｇｈ，Ｐ．）及びネルダー（Ｎｅｌｄｅｒ，Ｊ．Ａ．）の"Ｇｅｎｅｒａｌｉｚｅｄ　Ｌｉｎｅａｒ　Ｍｏｄｅｌｓ"（第2版、１９８９（Ｃｈａｐｍａｎ　ａｎｄ　Ｈａｌｌ　ｐｕｂ．）の第４章を参照のこと。

【００２６】当業者には明白であるように、プロセッサ１１は前述の方法を使用し、知的水準ファセットのような２値ではないファセットの各値を２値ファセットとみなすことによってこれらのファセットのための重み付けベクトルを生成することができる。即ち、非２値ファセットの各値に対して重み付けベクトルを生成する。

【００２７】好適な数（５０～５５）のキューを用いたロジスティック回帰を使用すると、オーバーフィッティング（ｏｖｅｒｆｉｔｔｉｎｇ）を生じる場合がある。更に、ロジスティック回帰は可変の相互作用のモデルを作らない。可変相互作用のモデリングを可能としてオーバーフィッティングを避けるために、ニューラルネットワークをステップ５４に使用して重み付けベクトルを生成し、性能を改良することができる。しかし、どちらのアプローチも本発明と矛盾せずステップ５４で使用することができる。

【００２８】後のテキストジャンルの自動識別を可能にするために、プロセッサ１１は選択されたファセットの各々に対する重み付けベクトルをメモリに記憶する。これが終了すると、学習は完了する。

【００２９】Ｄ．テキストジャンル及びファセットの自動識別

図４は、命令１００をフロー図で示している。命令１００を実行することで、プロセッサ11は、表面レベルのキュー、ファセットのセット及び重み付けベクトルを用いてマシン可読でタグ付けされていないテキスト１１のテキストジャンルを自動的に識別する。簡潔に説明すると、命令１００に従って、プロセッサ１１はまず、分類されるべきトークン化マシン可読テキストのキューベクトルを生成する。続いて、プロセッサ１１はキューベクトルとファセットに関連する重み付けベクトルとを使用して、各ファセットのテキストとの関連性を決定する。各ファセットのテキストとの関連性を決定した後、プロセッサ１１はテキストのジャンル（単数又は複数）を識別する。命令１００は、固体メモリ又はフロッピーディスクドライブ内に配置したフロッピーディスクに記憶させることができ、ＬＩＳＰ及びＣ＋＋を含むあらゆるコンピュータ言語で実現させることができる。

【００３０】選択されたトークン化マシン可読テキストのジャンルを識別するというユーザの要求に応答して、プロセッサ１１はステップ１０２に進む。このステップにおいて、プロセッサ１１はテキストのためのキューベクトルXを生成する。これは、選択されたテキスト内の、先に定義した表面レベルキューの各々に対する観測値を表している。前述したように、キューの定義に基づいてキュー値を決定する方法は当業者には明白であり、本明細書に詳しく説明する必要はない。次に、プロセッサ１１はステップ１０４に進み、選択されたテキストに関連するファセットを識別するプロセスを開始する。

【００３１】命令１００に従って、関連ファセットの識別は２値のファセットを用いて始まる。しかし、本発明と矛盾せず、非２値の値のファセットを用いて識別を始めてもよい。２値ファセットの評価は、プロセッサ１１がステップ１０４において１つのファセットを選択することによって始まる。

【００３２】次にプロセッサ１１は選択されたファセットに関連する重みベクトルβをメモリから検索し、ステップ１０２において生成されたキューベクトルXと重みベクトルβとを組み合わせる。プロセッサ１１は、これらの２つのベクトルを組み合わせ、選択されたファセットの分類されるテキストとの関連性のインジケータを生成するために多数の数学的アプローチを使用することができ、これらにはロジスティック回帰及びログ奇関数が含まれる。学習の際の使用とは反対に、プロセッサ１１はステップ１０６においてログ奇関数を解いてψを得る。ψは、ここでは選択されたファセットのテキストとの関連性を表す。ログ奇関数の答が０よりも大きい値を生じた場合、プロセッサ１１はファセットをテキストと関連性があるものとみなすが、本発明と矛盾せず関連性のカットオフ値として他の値を選択してもよい。

【００３３】１つの２値ファセットの関連性を決定した後、プロセッサ１１はステップ１０８に進み、他の２値ファセットの評価が必要であるか否かを確認する。必要であれば、プロセッサ11は分岐してステップ104に戻り、全ての２値ファセットを処理するまでステップ１０４、１０６及び１０８のループを実行することによって、１度に１つずつファセットの関連性の評価を続ける。２値ファセットの処理が終わると、プロセッサ１１はステップ１０８から分岐してステップ１１０に進み、非２値ファセットの関連性を決定するプロセスを始め

る。

【００３４】ここでもまた、プロセッサ１１はループを実行して非２値ファセットの関連性を決定する。各ファセット値を別個に評価しなくてはならないという点で、非２値ファセットの処理は２値ファセットのそれとは異なっている。従って、ステップ１１４を繰り返し実行することによって選択されたファセットの各値に対するログ奇関数の値を生成した後、プロセッサ１１はステップ１１８においてどのファセット値が最も関連しているかを決定しなくてはならない。プロセッサ１１は、スコアが最も高いファセット値を最も関連しているものとみなす。非２値ファセットの各々に対して適切なファセット値を決定した後、プロセッサ１１はステップ１２０からステップ１２２に進む。

【００３５】ステップ１２２において、プロセッサ１１は、関連性があると判断したファセットとファセット値によるテキストジャンルの定義とを用いて、選択されたテキストがどのテキストジャンルを表すかを識別する。これを実行する方法は当業者には明白であり、本明細書に詳しく説明する必要はない。この後、プロセッサ１１は、選択されたテキストに関連性があると判断されたテキストジャンル及びファセットを選択されたテキストに関連させる。ステップ１２２におけるテキストジャンルの決定は好ましいものであるが、これは任意である。何故なら、前述のように、ファセット分類はそのものが有用であるためテキストジャンルを定義しなくてもよいからである。

【００３６】Ｅ．テキストジャンル及びファセット分類のアプリケーション

自然言語分野及び情報検索分野は共に、テキストジャンル及びファセットの自動分類の多数のアプリケーション（用途）を提供する。自然言語では、自動テキスト分類はタガー及び翻訳において有用である。情報検索分野では、テキストジャンル分類は、文書の書式（フォーマット）の改訂及び自動要約の強化の際に検索フィルタ及びパラメータとして有用である。

【００３７】現行の意味タガー及び品詞タガーは共に、テキスト内のアイテムの頻度数に関する生統計を使用している。テキストジャンルに従ってテキストを自動的に分類し、テキストジャンルに従ってタガーに関連する確率を計算することにより、これらのタガーの性能を改良することができる。例えば、″ｓｏｒｅ″という単語が「怒った」という意味を有する確率又は″ｃｏｏｌ″という単語が「すばらしい」という意味を有する確率は、批評家の伝記においてよりも新聞のショートストーリーの映画の批評においてずっと高い。

【００３８】言語翻訳システム及び言語生成システムは共に、同義語のセット同士の区別をする。どの同義語のセットを選択すべきかを示す条件は複雑であり、調節が必要である。言語翻訳システムは、元の言語における単語の意味を認識し、標的言語における適切な同義語を識別しなくてはならない。これらの難点は、例えば同じ「スラング」のフランス語の単語を英語の同等の「スラング」に無条件に置き換えるなど、単に各言語のアイテムをラベル付けして言語間で系統的に翻訳するだけでは解決することができない。″Ｉｌ　ｃｈｅｒｃｈｅ　ｕｎ　ｂｏｕｌｏｔ″というフランス語の文は、１つの文脈では「彼は一夜興行（ｇｉｇ）を探している」と翻訳され、別の文脈では「彼は仕事を探している」と翻訳されうる。″Ｉｌ（ｒｅ）ｃｈｅｒｃｈｅ　ｕｎ　ｔｒａｖａｉｌ″という文は、「彼は仕事を探している」又は「彼は雇用を求めている」になる、などである。適切な選択は、ソースアイテムが得られるテキストのジャンルの分析に依存する。自動テキストジャンル分類は、言語翻訳システム及び言語生成システム双方の性能を改良することができる。何故ならば、この分類によって言語の種々のテキストジャンル及び種々のレジスター、従って多くの同義語セットのメンバ間の区別を認識することができるからである。このような同義語セットには以下が含まれる：″ｄｉｓｍｉｓｓ／ｆｉｒｅ／ｃａｎ″、″ｒａｔｈｅｒ／ｐｒｅｔｔｙ″、″ｗａｎｔ／ｗｉｓｈ″、″ｂｕｙ　ｉｔ／ｄｉｅ／ｄｅｃｅａｓｅ″、″ｗｈｅｅｌｓ／ｃａｒ／ａｕｔｏｍｏｂｉｌｅ″及び″ｇｉｇ／ｊｏｂ／ｐｏｓｉｔｉｏｎ″。

【００３９】多くの情報検索システムは同種のデータベースを用いて発達しており、これらの異種のデータベースに対する実行が不十分な傾向にある。自動テキストジャンル分類は、トピックベースの検索の出力に対するフィルタ又は独立した検索パラメータとして動作することにより、異種のデータベースを有する情報検索システムの性能を改良することができる。例えば、検索者はスーパーコライダーに関する新聞の論説を検索するが新聞の記事を除いて検索したい場合や、専門誌ではなく大衆雑誌においてＬＡＮＳに関する記事を検索したい場合がある。同様に、検索者は特定のテキストを用いて検索を開始し、そのテキストのジャンル及びトピックと類似する他のテキストを検索するように検索システムに要求する場合がある。情報検索システムは、トピックベースの検索の結果をランク付け又はクラスタリングする１つの方法として、ジャンル分類を使用することができる。

【００４０】また、自動ジャンル分類は文書の書式に関連する情報検索アプリケーション（用途）を有する。現在、多数の文書データベースが、電子テキストの外観に関する情報を含んでいる。例えば、マークアップ言語はインターネット上のデジタルテキストの書式を指定するために頻繁に使用されている。ハードコピー文書のＯＣＲも、大量の書式情報を含む電子文書を生成している。しかし、書式特徴の意味は、異種のデータベース内でジャンルによって様々でありうる。１つの例として、テキストにおいてボールドフェース及び通常のタイプを交互

に使用することを考えてみる。雑誌の記事では、この書式特徴は恐らくインタビューを示す。百科事典では、この同一の特徴は見出しと後に続くテキストを示す。マニュアルでは、この特徴は非常に重要であるか又はさほど重要ではない情報を示すために使用されうる。しかし、Ｗｉｒｅｄという雑誌では、この書式特徴は様々な記事を区別するために使用されている。自動テキストジャンル分類を使用して書式特徴の意味を決定することは、多数のアプリケーションにおいて有用である。このようにすることで、ユーザは見出し、要約及びタイトルなどの主な分野又は文書ドメインに検索を限定することができる。同様に、書式特徴の意味を決定することにより、自動文書要約、トピックのクラスタリング及び他の情報検索タスクの際に、非常に重要な文書ドメインとさほど重要ではない文書ドメインとの間の区別をつけることができる。また、書式特徴の意味を決定することにより、元の書式を保存することができないか又は望まれないいくつかの状況において、デジタル文書を新しい書式で表すことができる。例えば、いくつかの既存テキストを異なる書式のスタイルと組み合わせることによって新しい文書を生成する際は、均一な書式が望ましい。

【００４１】同様に、自動ジャンル分類は、書式設定されていないＡＳＣＩＩテキストの書式設定の態様を決定する際に有用である。

【００４２】テキストジャンルの自動分類は、自動文書要約に対して多数のアプリケーションを有する。第１に、いくつかの自動要約システムは、文を抽出すべきか否かを決定する際の特徴として、パラグラフ内の文の相対位置を使用している。しかし、文の特定の位置の有意性はジャンルによって様々である。新聞の記事の冒頭付近の文は、終わり付近の文よりも重要である可能性が高い。これは、法的決定及び雑誌のストーリーのような他のジャンルでは異なることが想定される。これらの相関関係は、自動ジャンル分類を用いて経験的に決定することができる。第２に、ジャンル分類により、要約されるテキストのジャンルに適した要約を作成することができる。読者が適切であると考える要約はジャンルによって異なるため、これは望ましいことである。自動要約システムは、前置きの文があるためにどこからテキストが始まるかを決定するのが困難である場合が多く、これは自動ジャンル分類の第３のアプリケーションを生じる。テキストに関連する前置きの文は、テキストのジャンルによって異なることが多い。

【図面の簡単な説明】

【図１】マシン可読テキストのテキストジャンルを自動的に決定するコンピュータシステムを示している。

【図２】ファセット値に従った表面キュー値の試行観察の表である表１を示している。

【図３】学習コーパスから重み付けベクトル値を生成する学習のためのフロー形式の命令を示している。

【図４】テキストジャンル及びファセットのマシン可読テキストとの関連を決定するフロー形式の命令を示している。

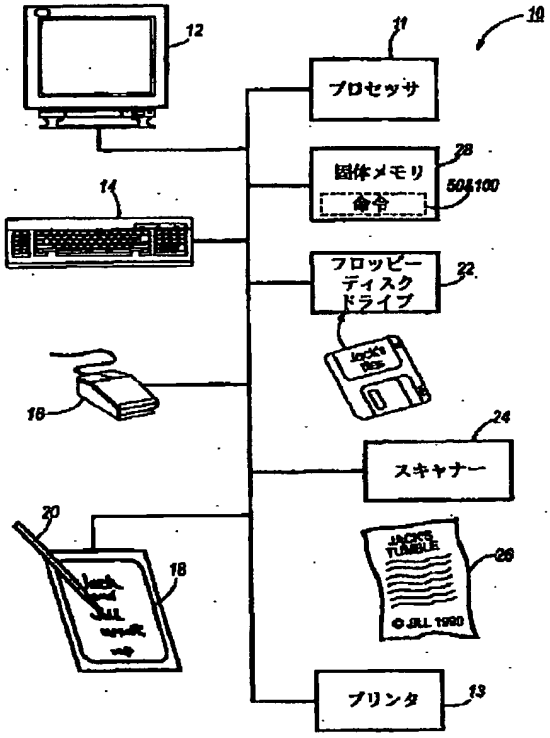
【図５】テキストジャンル又はファセット値に基づいた順序で検索結果をユーザに提示するためのフロー形式の命令を示している。

【図６】検索結果をコンピュータユーザに提示するためのフロー形式の命令を示している。

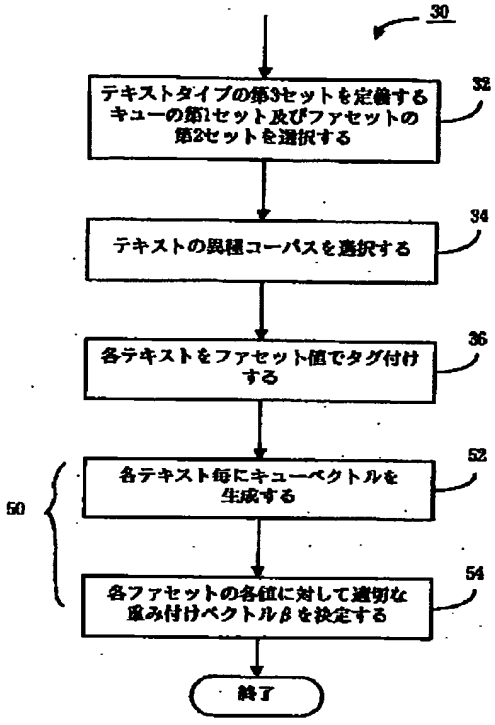
【符号の説明】

１０　コンピュータシステム
１１　プロセッサ
１２　モニタ
１３　プリンタ
１４　キーボード
１６　マウス
１８　電子タブレット
２０　スタイラス
２２　フロッピーディスクドライブ
２４　スキャナー
２６　テキスト
２８　固体メモリ
５０、１００　命令

【図1】



【図3】

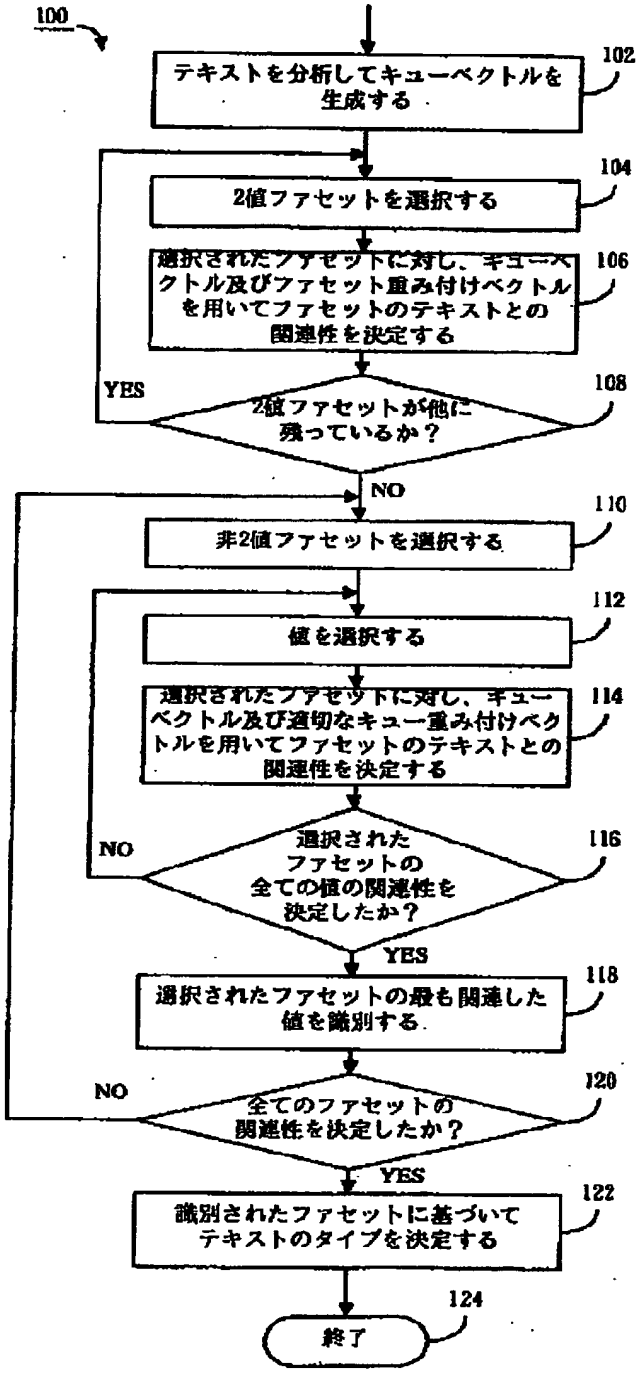


【図2】

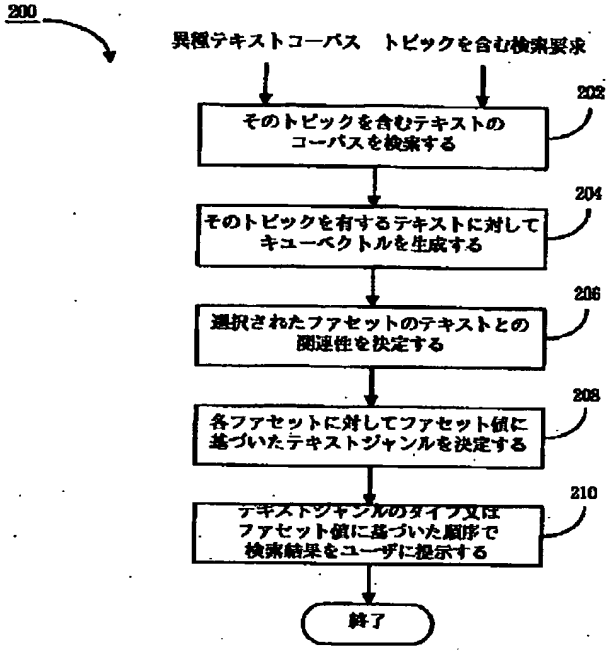
表1

キュータイプ	ファセットタイプ							
						知的水準		
	新聞の報道	雑誌の意見	法的	科学及び技術的	フィクション	平俗	中	高
平均値（文字／単語）／記事	4.94	4.85		4.81	4.25	4.83	4.80	4.93
コンマの平均数／記事	110.7	118.6		78.11	120.20	108.00	118.9	102.00
平均値（ダッシュ／文）／記事	1.22	1.21		0.87	0.90	1.20	1.38	1.43
平均値（コンマ／文）／記事	.09	0.1		0.03	0.05	0.14	0.09	
クエスチョンマークの平均数	0.6	5		0.22	0.50	0.83	2.50	1.75
平均値（クエスチョンマーク／文）／記事	0.1			0.0	0.06	0.01	0.04	0.02
ダッシュの平均数／記事	7.9	9.60		2.67	6.60	10.33	9.10	5.50
セミコロンの平均数／記事	1.20	3.20	4.78		4.80	3.67	5.60	7.75
平均値（セミコロン／文）／記事	0.01	0.03	0.05		0.04	0.04	0.06	0.11
"and"、"but"及び"so as"で始まる文／記事	3.5	7.7		1.0	6.7	6.7	9.1	4.0
副詞＋コンマで始まる文／記事	1.7	2.9		5.3	2.8	2.8	2.6	3.2

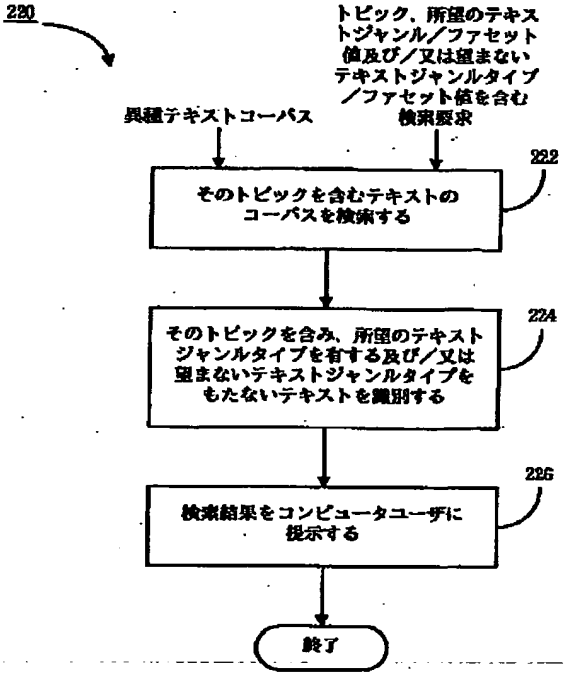
【図4】


100
テキストを分析してキューベクトルを生成する
102
2値ファセットを選択する
104
選択されたファセットに対し、キューベクトル及びファセット重み付けベクトルを用いてファセットのテキストとの関連性を決定する
106
YES
2値ファセットが他に残っているか？
108
NO
非2値ファセットを選択する
110
値を選択する
112
選択されたファセットに対し、キューベクトル及び適切なキュー重み付けベクトルを用いてファセットのテキストとの関連性を決定する
114
NO
選択されたファセットの全ての値の関連性を決定したか？
116
YES
選択されたファセットの最も関連した値を識別する
118
NO
全てのファセットの関連性を決定したか？
120
YES
識別されたファセットに基づいてテキストのタイプを決定する
122
終了
124

【図5】


200
異種テキストコーパス
トピックを含む検索要求
そのトピックを含むテキストのコーパスを検索する
202
そのトピックを有するテキストに対してキューベクトルを生成する
204
選択されたファセットのテキストとの関連性を決定する
206
各ファセットに対してファセット値に基づいたテキストジャンルを決定する
208
テキストジャンルのタイプ又はファセット値に基づいた順序で検索結果をユーザに提示する
210
終了

【図6】


220
異種テキストコーパス
トピック、所望のテキストジャンル／ファセット値及び／又は望まないテキストジャンルタイプ／ファセット値を含む検索要求
そのトピックを含むテキストのコーパスを検索する
222
そのトピックを含み、所望のテキストジャンルタイプを有する及び／又は望まないテキストジャンルタイプをもたないテキストを識別する
224
検索結果をコンピュータユーザに提示する
226
終了

フロントページの続き

(72)発明者　ハインリッチ　シェッツェ
アメリカ合衆国　94305　カリフォルニア州　スタンフォード　ベンチュラー　ホール　シーエスエルアイ（番地なし）
(72)発明者　ジャン　オー．ペダーセン
アメリカ合衆国　94555　カリフォルニア州　フレモント　ウェルマン　テラス　34398
(72)発明者　ブレット　エル．ケッセラー
アメリカ合衆国　94025　カリフォルニア州　メンロパーク　サンアントニオ　アベニュー　1508　アパートメント　エヌ
(72)発明者　グレゴリー　グレフェンステッテ
フランス国　デエレツ　サン　マルタン　38400　アベニュー　デ　ラ　ガロチェレ　21

【外国語明細書】

1 Title of Invention

ARTICLE AND METHOD OF AUTOMATICALLY DETERMINING TEXT GENRE USING SURFACE FEATURES OF UNTAGGED TEXTS

2 Claims

1. A processor implemented method of identifying a text genre of an untagged text in machine readable form without structurally analyzing the text, the processor implemented method comprising the steps of:
   a) generating a cue vector from the text, the cue vector representing occurrences in the text of a first set of nonstructural, surface cues; and
   b) determining whether the text is an instance of a first text genre using the cue vector and a weighting vector associated with the first text genre.

3 Detailed Description of Invention

Field of the Invention

The present invention relates to computational linguistics.

Background of the Invention

The word "genre" usually functions as a literary substitute for "kind of text." Text genre differs from the related concepts of text topic and document genre Text genre and text topic are not wholly independent. Distinct text genres like newspaper stories, novels and scientific articles tend to largely deal with different ranges of topics; however, topical commonalties within each of these text genres are very broad and abstract. Additionally, any extensive collection of texts relating to a single topic almost always includes works of more than one text genre so that the formal similarities between them are limited to the presence of lexical items. While text genre as a concept is

independent of document genre, the two genre types grow up in close historical association with dense functional interdependencies. For example, a single text genre may be associated with several document genres. A short story may appear in a magazine or anthology or a novel can be published serially in parts, reissued as a hard cover and later as a paper back. Similarly, a document genre like a newspaper may contain several text genres, like features, columns, advice-to-the-lovelorn, and crossword puzzles. These text genres might not read as they do if they did not appear in a newspaper, which licenses the use of context dependent words like "yesterday" and "local". By virtue of their close association, material features of document genres often signal text genre. For example, a newspaper may use one font for the headlines of "hard news" and another in the headlines of analysis; a periodical may signal its topical content via paper stock; business and personal letters can be distinguished based upon page lay out; and so on. It is because digitization eliminates these material clues as to text and document genres that it is often difficult to retrieve relevant texts from heterogeneous digital text collections.

The boundaries between textual genres mirror the divisions of social life into distinct roles and activities - between public and private, generalist and specialist, work and recreation, etc. Genres provide the context that makes documents interpretable, and for this reason genre, no less than content, shapes the user's conception of relevance. For example, a researcher seeking information about supercolliders or Napoleon will care as much about text genre as content - she will want to know not just what the source says, but whether that source appears in a scholarly journal or in a popular magazine.

Until recently work on information retrieval and text classification has focused almost exclusively on the identification of topic, rather than on text genre. Two reasons explain this neglect. First, the traditional print-based document world did not perceive a need for genre classification because in this world genres are clearly marked, either intrinsically or by institutional and

contextual features. A scientist looking in a library for an article about cold fusion need not worry about how to restrict his search to journal articles, which are catalogued and shelved so as to keep them distinct from popular science magazines. Second, early information retrieval work with on-line text databases focused on small, relatively homogeneous databases in which text genre was externally controlled, like encyclopedia or newspaper databases. The creation of large, heterogeneous, text databases, in which the lines between text genres are often unmarked, highlights the importance of genre classification of texts. Topic-based search tools alone cannot adequately winnow the domain of a reader's interest when searching a large heterogeneous database.

Applications of genre classification are not limited to the field of information retrieval. Several linguistic technologies could also profit from its application. Both automatic part of sentence taggers and sense taggers could benefit from genre classification because it is well known that the distribution of word senses varies enormously according to genre.

Discussions of literary classification stretch back to Aristotle. The literature on genre is rich with classificatory schemes and systems, some of which might be analyzed as simple attribute systems. These discussions tend to be vague and to focus exclusively on literary forms like the eclogue or the novel, and, to a lesser extent, on paraliterary forms like the newspaper crime report or the love letter. Classification discussions tend to ignore unliterary textual types such as annual reports, Email communications and scientific abstracts. Moreover none of these discussions make an effort to tie the abstract dimensions along which genres are distinguished to any formal features of the texts.

The only linguistic research specifically concerned with quantificational methods of genre classification of texts is that of Douglas Biber. His work includes: Spoken and Written Textual Dimensions in English: Resolving the

Contradictory Findings, Language, 62(2):384-413, 1986; Variation Across Speech and Writing, Cambridge University Press, 1988; The Multidimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Finding, Computers in the Humanities, 26(5-6):331-347, 1992; Using Register-Diversified Corpora for General Language Studies, in Using Large Corpora, pp. 179-202 (Susan Armstrong ed.) (1994); and with Edward Finegan, Drift and the Evolution of English Style: A History of Three Genres, Language, 65(1):93-124, 1989. Biber's work is descriptive, aimed at differentiating text genres functionally according to the types of linguistic features that each tends to exploit. He begins with a corpus that has been hand-divided into a number of distinct genres, such as "academic prose" and "general fiction " He then ranks these genres along several textual "dimensions" or factors, typically three or five. Biber individuates his factors by applying factor analysis to a set of linguistic features, most of them syntactic or lexical. These factors include, for example, past-tense verbs, past participial clauses and "wh-" questions. He then assigns to his factors general meanings or functions by abstracting over the discourse functions that linguists have applied assigned to the individual components of each factor; e g., as an "informative vs. involved" dimension, a "narrative vs. non-narrative" dimension, and so on. Note that these factors are not individuated according to their usefulness in classifying individual texts according to genre. A score that any text receives on a given factor or set of factors may not be greatly informative as its genre because there is considerable overlap between genres with regard to any individual factor.

Jussi Karlgren and Douglass Cutting describe their effort to apply some of Biber's results to automatic categorization of genre in Recognizing Text Genres with Simple Metric Using Discriminant Analysis, in Proceedings of Coling '94, Volume II, pp. 1071-1075, Aug. 1994. They too begin with a corpus of hand classified texts, the Brown corpus. The people who organized the Brown

corpus describe their classifications as generic, but the fit between the texts and the genres a sophisticated reader would recognize is only approximate. Karlgren and Cutting use either lexical or distributional features - the lexical features include first-person pronoun count and present-tense verb count, while the distributional features include long-word count and character per word average. They do not use punctuational or character level features. Using discriminant analysis, the authors classify the texts into various numbers of categories. When Karlgren and Cutting used a number of functions equal to the number of categories assigned by hand, the fit between the automatically derived and hand-classified categories is 51.6%. They improved performance by reducing the number of functions and reconfiguring the categories of the corpus. Karlgren and Cutting observe that it is not clear that such methods will be useful for information retrieval purposes, stating: "The problem with using automatically derived categories is that even if they are in a sense real, meaning that they are supported by the data, they may be difficult to explain for the unenthusiastic layman if the aim is to use the technique in retrieval tools." Additionally, it is not clear to what extent the idiosyncratic "genres" of the Brown corpus coincide with the categories that users find relevant for information retrieval tasks.

Geoffrey Nunberg and Patrizia Violi suggest that genre recognition will be important for information retrieval and natural language processing tasks in Text Form and Genre in Proceedings of QED'92, pp. 118-122. October 1992. These authors propose that text genre can be treated in terms of attributes, rather than classes; however, they offer no concrete proposal as to how identification can be accomplished.

Summary of the Invention

The method of the present invention for automatically identifying the genre of a machine readable, untagged, text provides these and other advantages. Briefly described, the processor implemented method begins by

generating a cue vector from the text, which represents occurrences in the text of a first set of nonstructural, surface cues, which are easily computable. Afterward, the processor determines whether the text is an instance of a first text genre using the cue vector and a weighting vector associated with the first text genre.

Detailed Description of the Preferred Embodiments

Figure 1 illustrates in block diagram form computer system 10 in which the present method is implemented by executing instructions 100. The present method alters the operation of computer system 10, allowing it to automatically determine the text genre of an untagged text presented to it in machine readable form. Instructions 100 enable text genre classification to occur without structural analysis of the text, word stemming or part of speech tagging. Instructions 100 rely upon new surface-level cues, or features, which can be computed more quickly than structurally based features. Briefly described, according to instructions 100, computer system 10 analyzes the text to determine the number of occurrences of each surface cue within the text generates a cue vector. Computer system 10 then determines whether the text is an instance of a particular text genre and/or facet using the cue vector and a weighting vector associated with the particular text genre and/or facet. Instructions 100 will be described in detail with respect to Figure 4. Computer system 10 determines the appropriate weighting vector for each text genre and/or facet using training instructions 50, which will be described in detail with respect to Figure 3.

A. A Computer System for Automatically Determining Text Genre

Prior to a more detailed discussion of instructions 50 and 100 consider computer system 10, which executes those instructions, Illustrated in Figure 1, computer system 10 includes monitor 12 for visually displaying information to a computer user. Computer system 10 also outputs information to the

computer user via printer 13. Computer system 10 provides the computer user multiple avenues to input data. Keyboard 14 allows the computer user to input data to computer system 10 by typing. By moving mouse 16 the computer user is able to move a pointer displayed on monitor 12. The computer user may also input information to computer system 10 by writing on electronic tablet 18 with a stylus 20 or pen. Alternately, the computer user can input data stored on a magnetic medium, such as a floppy disk, by inserting the disk into floppy disk drive 22. Scanner 24 allows the computer user to generate machine readable versions, e.g. ASCII, of hard copy documents.

Processor 11 controls and coordinates the operations of computer system 10 to execute the commands of the computer user. Processor 11 determines and takes the appropriate action in response to each user command by executing instructions, which like instructions 50 and 100, are stored electronically in memory, either memory 28 or on a floppy disk within disk drive. Typically, operating instructions for processor 11 are stored in solid state memory, allowing frequent and rapid access to the instructions. Semiconductor logic devices that can be used to realize memory include read only memories (ROM), random access memories (RAM), dynamic random access memories (DRAM), programmable read only memories (PROM), erasable programmable read only memories (EPROM), and electrically erasable programmable read only memories (EEPROM), such as flash memories.

B. Text Genres, Facets and Cues

According to instructions 50 and 100, computer system 10 determines the text genre of a tokenized, machine readable text that has not been structurally analyzed, stemmed, parsed, nor tagged for sense or parts of speech. As used herein, a "text genre" is any widely recognized class of texts defined by some common communicative purpose or other functional traits, provided that the function is connected to some formal cues or commonalties that are not the direct consequences of the immediate topic that the texts address. Wide

recognition of a class of texts enables the public to interpret the texts of the class using a characteristic set of principles of interpretation. As used herein, text genre applies only to sentential genres; that is, applies only to genres that communicate primarily via sentences and sentence like strings that make use of the full repertory of text-category indicators like punctuation marks, paragraphs, and the like. Thus, according to the present invention airline schedules, stock tables and comic strips are not recognized as text genres. Nor does the present invention recognize genres of spoken discourse as text genres. Preferably, the class defined by a text genre should be extensible. Thus, according to the present invention the class of novels written by Jane Austen is not a preferred text genre because the class is not extensible.

The methods of instructions 50 and 100 treat text genres as a bundle of facets, each of which is associated with a characteristic set of computable linguistic properties, called cues or features, which are observable from the formal, surface level, features of texts. Using these cues, each facet distinguishes a class of texts that answer to certain practical interests. Facets tend to identify text genre indirectly because one facet can be relevant to multiple genres. Because any text genre can be defined as a particular cluster of facets the present method allows identification of text genres and supergenres with the same accuracy as other approaches, but with the advantage of easily allowing the addition of new, previously unencountered text genres.

Rather than attempting to further define the concept of facets, consider a number of illustrative examples. The audience facet distinguishes between texts that have been broadcast and those whose distribution was directed to a more limited audience. The length facet distinguishes between short and long texts. Distinctions between texts that were authored by organizations or anonymously and individuals are represented by the author facet. List below are other facets and their values, when those values are not obvious. Note

facets need not be binary valued.

Facet Name	Possible Values	
1. Date	Dated/Undated	
2. Narrative	Yes/No	
3. Suasive(Argumentative)/Descriptive(Informative)		
4. Fiction/Nonfiction		
5. Legal	Yes/No	
6. Science & Technical	Yes/No	
7. Brow	Popular	Yes/No
	Middle	Yes/No
	High	Yes/No

Other facets can be defined and added to those listed above consistent with the present invention. Not all facets need be used to define a text genre; indeed, a text genre could be defined by a single facet. Listed below are but a few examples of conventionally recognized text genres that can be defined using the facets and values described.

1. Press Reports

a.	Audience	Broadcast
b.	Date	Dated
c.	Suasive	Descriptive
d.	Narrative	Yes
e.	Fiction	No
f.	Brow	Popular
g.	Author	Unsigned
h.	Science&Technical	No
i.	Legal	No

2. Editorial Opinions

a.	Audience	Broadcast
b.	Date	Dated
c.	Suasive	Yes
d.	Narrative	Yes
e.	Fiction	No
f.	Brow	Popular
g.	Authorship	Signed
h.	Science&Techinical	No
i.	Legal	No

3. Market Analysis

a.	Audience	Broadcast
b.	Date	Dated
c.	Suasive	Descriptive
d.	Narrative	No
e.	Fiction	No
f.	Brow	High
g.	Authorship	Organizational
h.	Science and Technical	Yes
i.	Legal	No

4. Email

a.	Audience	Directed
b.	Date	Dated
c.	Fiction	No
d.	Brow	Popular
e.	Authorship	Signed

Just as text genres decompose into a group of facets, so do facets decompose into surface level cues according to the present methods. The surface level cues of the present invention differ from prior features because

they can be computed using tokenized ASCII text without doing any structural analysis, such as word stemming, parsing or sense or part of speech tagging. For the most part, it is the frequency of occurrence of these surface level cues within a text that is relevant to the present methods. Several types of surface level or formal cues can be defined, including, but not limited to: numerical/statistical, punctuational, constructional, formulae, lexical and deviation. Formulae type cues are collocations or fixed expressions that are conventionally associated with a particular text genre. For example, fairy tales begin with "Once upon a time" and Marian hymns begin with "Hail Mary." Other formulae announce legal documents, licensing agreements and the like. Lexical type cues are directed to the frequency of certain lexical items that can signal a text genre. For example, the use of formal terms of address like "Mr., Mrs. and Ms." are associated with articles in the New York Times; and the use of words like "yesterday" and "local" frequently occur in newspaper reports. Additionally, the use of a phrase like "it's pretty much a snap" indicate that a text is not part of an encyclopedia article, for example. The use of some lexical items is warranted by the topical and rhetorical commonalties of some text genres. While constructional features are known in the prior art, computation of most of them requires tagged or fully parsed text. Two new surface level constructional cues are defined according to the present invention which are string recognizable. Punctuational type cues are counts of punctuational features within a text. This type of cue has not been used previously; however, they can serve as a useful indicator of text genre because they are at once significant and very frequent. For example, a high question mark count may indicate that a text attempts to persuade its audience. In contrast to most other cue types, which measure the frequency of surface level features within a particular text, deviation type cues relate to deviations in unit size. For example, deviation cues can be used to track variations in sentence and paragraph length, features that may vary according to text genre. Cue types

have been described merely to suggest the kinds of surface level features that can be measured to signal text features; characterization of cue type is not important to the present invention. The number of cues that can be defined is theoretically unlimited. Just a few of the possible cues are listed below for illustrative purposes.

A. Punctuational Cues
   1. Log (comma count +1)
   2. Mean (commas/sentences)/article
   3. Mean (dashes/sentences)/article
   4. Log (question mark count +1)
   5. Mean (questions/sentences)/article
   6. Log (dash count + 1)
   7. Log (semicolon count + 1)

B. String Recognizable Constructional Cues
   1. Sentences starting w/ "and" "but" and "so" per article
   2. Sentences starting w/adverb + comma/article

C. Formulae Cues
   1. "Once upon a time..."

D. Lexical Cues (Token counts only are taken unless otherwise indicated)
   1. Abbreviations for "Mr., Mrs." etc.
   2. Acronyms
   3. Modal auxiliaries
   4. Forms of the verb "be"
   5. Calendar - days of the week, months

6 , 7. Capital - non-sentence initial words that are capitalized

Type and Token counts

8. Number of characters

9,10. Contractions

Type and Token counts

11,12. Words that end in "ed"

Type and Token counts

13. Mathematical Formula

14. Forms of the verb "have"

15. 16 Hyphenated words

Type and token counts

17,18. Polysyllabic words

Type and token counts

19. The word "it"

20,21. Latinate prefixes and suffixes

Type and token counts

22,23. Words more than 6 letters

Type and token counts

24,25. Words more than 10 letters

Type and token counts

26,27. Three + word phrases

Type and token counts

28,29. Polysyllabic words ending in "ly"

Type and token counts

30. Overt negatives

31,32. Words containing at least one digit

Type and token counts

33. Left parentheses

34,35 Prepositions

Type and token counts

36. First person singular pronouns
37. First person plural pronouns
38. Pairs of quotation marks
39. Roman Numerals
40. Instances of "that"
41. Instances of "which"
42. Second person plural pronouns

F. Deviation Cues

1. standard deviation of sentence length in words
2. standard deviation of word length in characters
3. standard deviation of length of text segments between punctuation marks in words
4. Mean (characters/words) per article

The result of a preliminary trial with a corpus of approximately four hundred texts, Table I of Figure 2 illustrates how some surface level cues can vary according to face/text genre. (This trial treated some text genres as a single facet, rather than decomposing the text genres as described above. Both approaches are consistent with the present invention. As stated previously, a text genre may be defined by a single facet.) For example, within this corpus press reports included only 1.2 semicolons per article, while legal documents included 4.78. Similarly, the number of dashes per text differed among press reports, editorial opinions and fiction.

What weight should be given to different cue values? Or, stated another way, how strongly correlative is a cue value, or set of cue values, of a particular facet or text genre? In contrast to the decomposition of text genres into facet values, which is a matter of human judgment, answering this question is not. Determining the weight accorded to each cue according to facet requires

training, which is described below with respect to Figure 3.

C. Training to Determine Cue Weights

Figure 3 illustrates in flow diagram form training method 30 for determining cue weights for each cue, Training method 30 is not entirely automatic; steps 32, 34 and 36 are manually executed while those of instructions 50 are processor implemented. Instructions 50 may be stored in solid state memory or on a floppy disk placed within floppy disk drive and may be realized in any computer language, including LISP and C++.

Training method 30 begins with the selection of a set of cues and another set of facets, which can be used to define a set of widely recognized text genres. Preferably, about 50 to 55 surface level cues are selected during step 32, although a lesser or greater number can be used consistent with the present invention. Selection of a number of lexical and punctuational type surface level cues is also preferred. The user may incorporate all of the surface level cues into each facet defined, although this is not necessary. While any number of facets can be defined and selected during step 32, the user must define some number of them. In contrast, the user need not define text genres at this point because facets by themselves are useful in a number of applications, as will be discussed below. Afterward, during step 34 the user selects a heterogeneous corpus of texts. Preferably the selected corpus includes about 20 instances of each of the selected text genres or facets, if text genres have not been defined. If not already in digital or machine readable form, typically ASCII, then the corpus must be converted and tokenized before proceeding to instructions 50. Having selected facets, surface level cues and a heterogeneous corpus, during step 36 the user associates machine readable facet values with each of the texts of the corpus. Afterward, the user turns the remaining training tasks over to computer system 10.

Instructions 50 begin with step 52, during which processor 11 generates

a cue vector, X, for each text of the corpus. The cue vector is a multi-dimensional vector having a value for each of the selected cues. Processor 11 determines the value for each cue based upon the relevant surface level features observed within a particular text. Methods of determining cue values given definitions of the selected cues will be obvious to those of ordinary skill and therefore will not be described in detail herein. Because these methods do not require structural analysis or tagging of the texts, processor 11 expends relatively little computational effort in determining cue values during step 52.

Processor 11 determines the weighting that should be given to each cue according to facet value during step 54. In other words, during step 54 processor 11 generates a weighting vector, β, for each facet. Like the cue vector, X, the weighting vector, β, is a multidimensional vector having a value for each of the selected cues. A number of mathematical approaches can be used to generate weighting vectors from the cue vectors for the corpus, including logistic regression. Using logistic regression, processor 11 divides the cue vectors generated during step 52 into sets of identical cue vectors. Next for each binary valued facet, processor 11 solves a log odds function for each set of identical cue vectors. The log odds function, $g(\phi)$, is expressed as:

$$g(\phi) = \log(\phi/1-\phi) = X\beta;$$

where: ϕ is the proportion of vectors for which the facet value is true;
$1-\phi$ is the proportion of vectors in the set for which the facet value is false.

The processor 11 is able to determine the values of ϕ and $1-\phi$ because earlier tagging of facet values indicates the number of texts having each facet value within each set of texts having identical cue vectors. Thus, processor 11 can determine the values of weighting vector β for each binary valued facet by solving the system of simultaneous equations defined by all the sets of

identical cue vectors, the known values of ϕ,1-ϕ) and the cue vector values. Logistic regression is well known and will not be described in greater detail here. For a more detailed discussion of logistic regression, see Chapter 4 of McCullagh, P. and Nelder, J. A., Generalized Linear Models, 2d Ed , 1989 (Chapman and Hall pub), incorporated herein by reference.

Processor 11 can use the method just described to generate weighting vectors for facets that are not binary valued, like the Brow facet, by treating each value of the facet as a binary valued facet, as will be obvious to those of ordinary skill. In other words, a weighting vector is generated for each value of a non-binary valued facet.

Using logistic regression with as large a number of cues as preferred, 50-55, may lead to overfitting. Further, logistic regression does not model variable interactions. To allow modeling of variable interactions and avoid overfitting, neural networks can be used during step 54 to generate the weighting vectors and may improve performance. However, either approach may be used during step 54 consistent with the present invention.

To enable future automatic identification of text genre, processor 11 stores in memory the weighting vectors for each of the selected facets. That done, training is complete.

D. Automatically Identifying Text Genre and Facets

Figure 4 illustrates in flow diagram form instructions 100. By executing instructions 100, processor 11 automatically identifies the text genre of a machine readable, untagged, text 11 using set of surface level cues a set of facets and weighting vectors. Briefly described, according to instructions 100, processor 11 first generates a cue vector for the tokenized machine readable text to be classified. Subsequently, processor 11 determines the relevancy of each facet to the text using the cue vector and a weighting vector associated with the facet. After determining the relevancy of each facet to the text, processor 11 identifies the genre or genres of the text. Instructions 100 may be

stored in solid state memory or on a floppy disk placed within floppy disk drive and may be realized in any computer language, including LISP and C++.

In response to a user request to identify the genre of a selected tokenized, machine readable text, processor 11 advances to step 102. During that step, processor 11 generates for the text a cue vector, X, which represents the observed values within the selected text for each of the previously defined surface level cues. As discussed previously, methods of determining cue values given cue definitions will be obvious to those of ordinary skill and need not be discussed in detail here Processor 11 then advances to step 104 to begin the process of identifying the facets relevant to the selected text.

According to instructions 100, identification of relevant facets begins with the binary valued facets; however, consistent with the present invention identification may also begin with the non-binary valued facets. Evaluation of the binary valued facets begins with processor 11 selecting one during step 104.

Processor 11 then retrieves from memory the weight vector, θ, associated with the selected facet and combines it with the cue vector, X, generated during step 102. Processor 11 may use a number of mathematical approaches to combine these two vectors to produce an indicator of the relevance of the selected facet to the text being classified, including logistic regression and the log odds function. In contrast to its use during training, during step 106 processor 11 solves the log odds function to find ϕ, which now represents the relevance of the selected facet to the text. Processor 11 regards a facet as relevant to a text if solution of the log odds function produces a value greater than O, although other values can be chosen as a cut-off for relevancy consistent with the present invention.

Having determined the relevancy of one binary valued facet, processor 11 advances to step 108 to ascertain whether other binary-valued facets require evaluation. If so, processor 11 branches back up to step 104 and

continues evaluating the relevancy of facets, one at a time, by executing the loop of steps 104, 106 and 108 until every binary-valued facet has been considered. When that occurs, processor 11 branches from step 108 to step 110 to begin the process of determining the relevancy of the non-binary valued facets.

Processor 11 also executes a loop to determine the relevance of the non-binary valued facets. Treatment of the non-binary valued facets differs from that of binary valued facets in that the relevance of each facet value must be evaluated separately. Thus, after generating a value of the log odds function for each value of the selected facet by repeatedly executing step 114, processor 11 must decide which facet value is most relevant during step 118. Processor 11 regards the highest scoring facet value as the most relevant. After determining the appropriate facet value for each of the non-binary valued facets, processor 11 advances to step 122 from step 120.

During step 122 processor 11 identifies which text genres the selected text represents using the facets determined to be relevant and the text genre definitions in terms of facet values. Methods of doing so are obvious to those of ordinary skill and need not be described in detail herein. Afterward, processor 11 associates with the selected text. the text genres and facets determined to be relevant to the selected text. While preferred, determination of text genres during step 122 is optional because, as noted previously, text genres need not be defined because facet classifications are useful by themselves.

E. Applications for Text Genre and Facet Classification

The fields of natural language and information retrieval both present a number of applications for automatic classification of text genre and facets. Within natural language, automatic text classification will be useful with taggers and translation. Within the information retrieval field, text genre classification will be useful as a search filter and parameter, in revising document format and enhancing automatic summarization.

Present sense taggers and part of speech taggers both use raw statistics about the frequency of items within a text. The performance of these taggers can be improved by automatically classifying texts according to their text genres and computing probabilities relevant to the taggers according to text genre. For example, the probability that "sore" will have the sense of "angry" or that "cool" will have the sense of "first-rate" is much greater in a newspaper movie review of a short story than in a critical biography.

Both language translation systems and language generation systems distinguish between synonym sets. The conditions indicating which synonym of a set to select are complex and must be accommodated. Language translation system must recognize both the sense of a word in the original language and then identify an appropriate synonym in the target language. These difficulties cannot be resolved simply by labeling the items in each language and translating systematically between them; e.g., by categorically substituting the same "slang" English word for its "slang" equivalent in French. In one context the French sentence "ll cherche un boulot" might be translated by "He's looking for a gig," in another context by "He's looking for a job ". The sentence "ll (re)cherche un travail" might be either "He's looking for a job" or "He's seeking employment," and so on. Making the appropriate choice depends on an analysis of the genre of the text from which a source item derives. Automatic text genre classification can improve the performance of both language translation systems and language generation systems. It can do so because it allows recognition of different text genres and of different registers of a language, and, thus, distinctions between members of many synonym sets. Such synonym sets include: "dismiss/fire/can," "rather/pretty," "want/wish," "buy it/die/decease," "wheels/car/automobile" and "gig/job/position".

Most information retrieval system have been developed using homogeneous databases and they tend to perform poorly on heterogeneous databases. Automatic text genre classification can improve the performance of

information retrieval systems with heterogeneous databases by acting as a filter on the output of topic-based searches or as an independent search parameter. For example, a searcher might search for newspaper editorials on a supercollider, but exclude newspaper articles, or search for articles on LANS in general magazines but not technical journals. Analogously, a searcher might start with a particular text and ask the search system to retrieve other texts similar to it as to genre, as well as topic. Information retrieval systems could use genre classification as a way of ranking or clustering the results of a topic based search.

Automatic genre classification will also have information retrieval applications relating to document format. A great many document databases now include information about the appearance of the electronic texts they contain. For example, mark-up languages are frequently used to specify the format of digital texts on the Internet. OCR of hardcopy documents also produces electronic documents including a great deal of format information. However, the meaning of format features can vary within a heterogeneous database according to genre. As an example, consider the alternating use of boldface and normal type within a text. Within a magazine article this format feature likely indicates an interview; within an encyclopedia this same feature denotes headings and subsequent text; within a manual this feature may be used to indicate information of greater or lesser importance; or still yet, within the magazine Wired this format feature is used to distinguish different articles. Using automatic text genre classification to determine the meaning of format features would be useful in a number of applications. Doing so enables users to constrain their searches to major fields or document domains, like headings, summaries, and titles. Analogously, determining the meaning of format features enables discriminating between document domains of greater and lesser importance during automatic document summarization, topic clustering and other information retrieval tasks. Determining the meaning of format

features also enables the representation of digital documents in a new format. In a number of situations preservation of original format is impossible or undesirable. For example, a uniform format may be desired when generating a new document by combining several existing texts with different format styles.

In a similar vein, automatic genre classification is useful when determining how to format an unformatted ASCII text.

Automatic classification of text genre has a number of applications to automatic document summarization. First, some automatic summarizers use the relative position of a sentence within a paragraph as a feature in determining whether the sentence should be extracted. However, the significance of a particular sentence position varies according to genre. Sentences near the beginning of newspaper articles are more likely to be significant than those near the end. One assumes this is not the case for other genres like legal decisions and magazine stories. These correlations could be determined empirically using automatic genre classification. Second, genre classification allows tailoring of summaries according to the genre of the summarized text, which is desirable because what readers consider an adequate summary varies according to genre. Automatic summarizers frequently have difficulty determining where a text begins because of prefatory material, leading to a third application for automatic genre classification. Frequently, prefatory material associated with texts varies according to text genre.

4 Brief Description of the Drawings

Figure 1 illustrates a computer system for automatically determining the text genre of machine readable texts.

Figure 2 illustrates Table 1, a table of trial observations of surface cue values according to facet value.

Figure 3 illustrates in flow diagram form instructions for training to

generate weighting vectors values from a training corpus.

Figure 4 illustrates in flow diagram form instructions for determining the relevance of text genres and facets to a machine readable text.

Figure 5 illustrates in flow diagram form instruction for presenting search results to the user in an order based upon text genre type or facet values.

Figure 6 illustrates in flow diagram form instructions for presenting search results to the computer user.

FIG.1

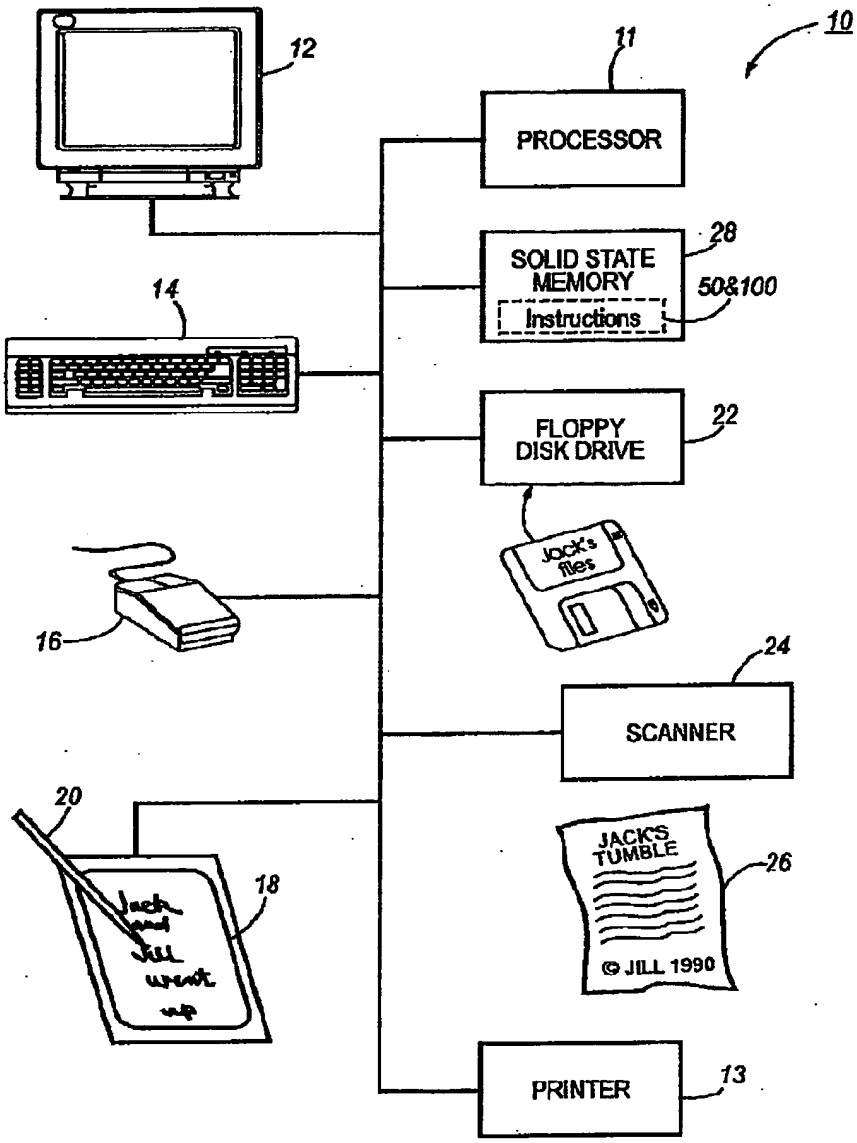

10
11
PROCESSOR
12
28
SOLID STATE MEMORY
Instructions
50&100
14
22
FLOPPY DISK DRIVE
Jock's files
16
24
SCANNER
20
18
JACKS TUMBLE
© JILL 1990
26
PRINTER
13

FIG.2

Table 1

Cue Type	Facet Type							
						Brow		
	Press Report	Editorial Opinion	Legal	Science & Technical	Fiction	Popular	Middle	High
Mean (Chars/words)/per article	4.94	4.85		4.81	4.25	4.63	4.80	4.93
Mean commas/article	110.7	119.6		78.11	120.20	109.00	118.9	102.00
Mean (commas/sentences)/article	1.22	1.21		0.87	0.90	1.2	1.36	1.43
Mean (dashes/sentences)/ article	0.09	0.1		0.03	0.05	0.14	0.09	
Mean question marks	0.6	5		0.22	8.50	0.83	2.90	1.75
Mean (questions/sentences)/article	0.1			0.0	0.06	0.01	0.04	0.02
Mean dashes/article	7.9	9.60		2.67	6.60	13.33	8.10	5.50
Mean semicolons/article	1.2	3.2	4.78		4.80	3.67	5.80	7.75
Mean (semicolons/sentences)/ article	0.01	0.03	0.05		0.04	0.04	0.06	0.11
Sentences starting w/"and", "but", and "so" per article	3.5	7.7		1.0	6.7	6.7	9.1	4.0
Sentence starting w/ adverb + comma /article	1.7	2.9		5.3	2.8	2.3	2.6	3.2

特開平11-120206

FIG.3

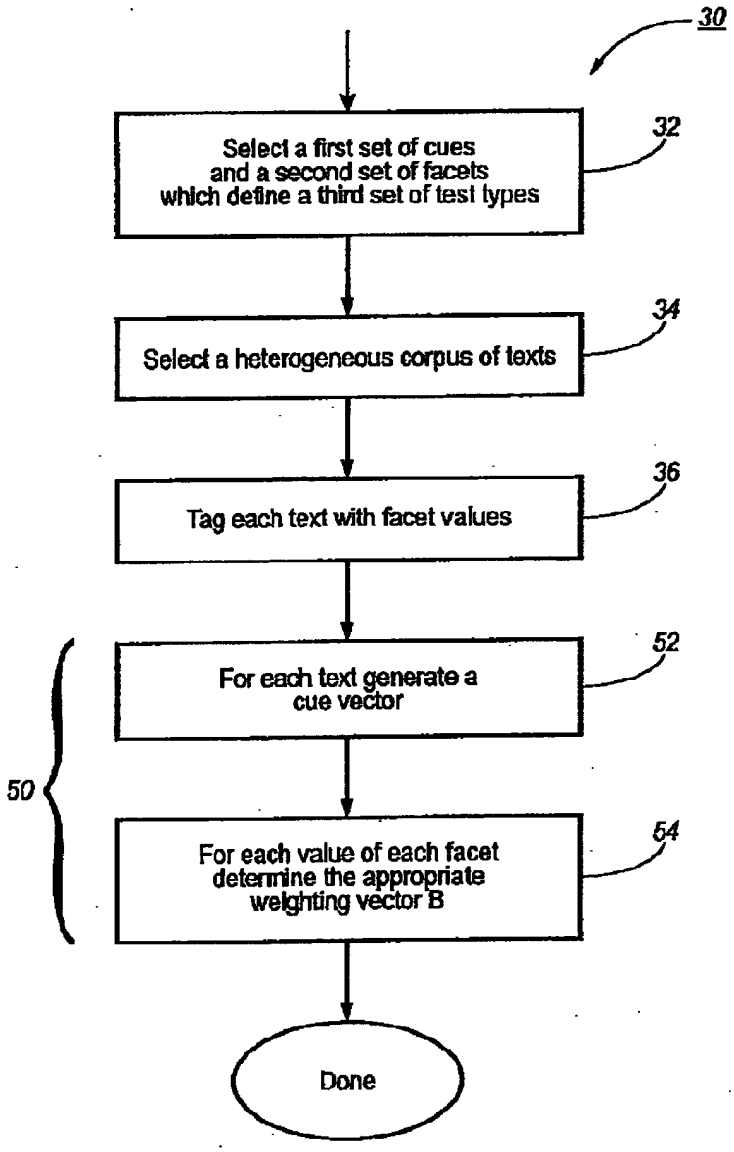

30
Select a first set of cues and a second set of facets which define a third set of test types
32
Select a heterogeneous corpus of texts
34
Tag each text with facet values
36
For each text generate a cue vector
52
50
For each value of each facet determine the appropriate weighting vector B
54
Done

FIG.4

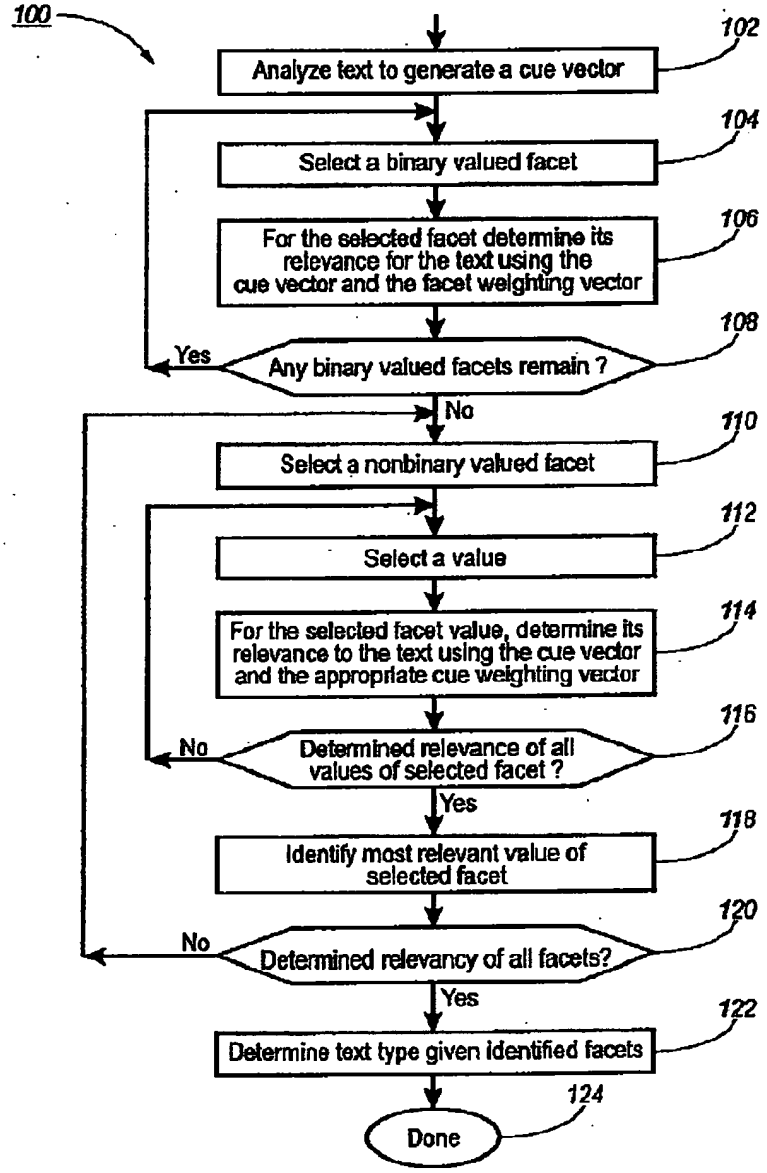

100
Analyze text to generate a cue vector
102
Select a binary valued facet
104
For the selected facet determine its relevance for the text using the cue vector and the facet weighting vector
106
Any binary valued facets remain ?
108
Yes
No
Select a nonbinary valued facet
110
Select a value
112
For the selected facet value, determine its relevance to the text using the cue vector and the appropriate cue weighting vector
114
Determined relevance of all values of selected facet ?
116
No
Yes
Identify most relevant value of selected facet
118
Determined relevancy of all facets?
120
No
Yes
Determine text type given identified facets
122
Done
124

FIG.5

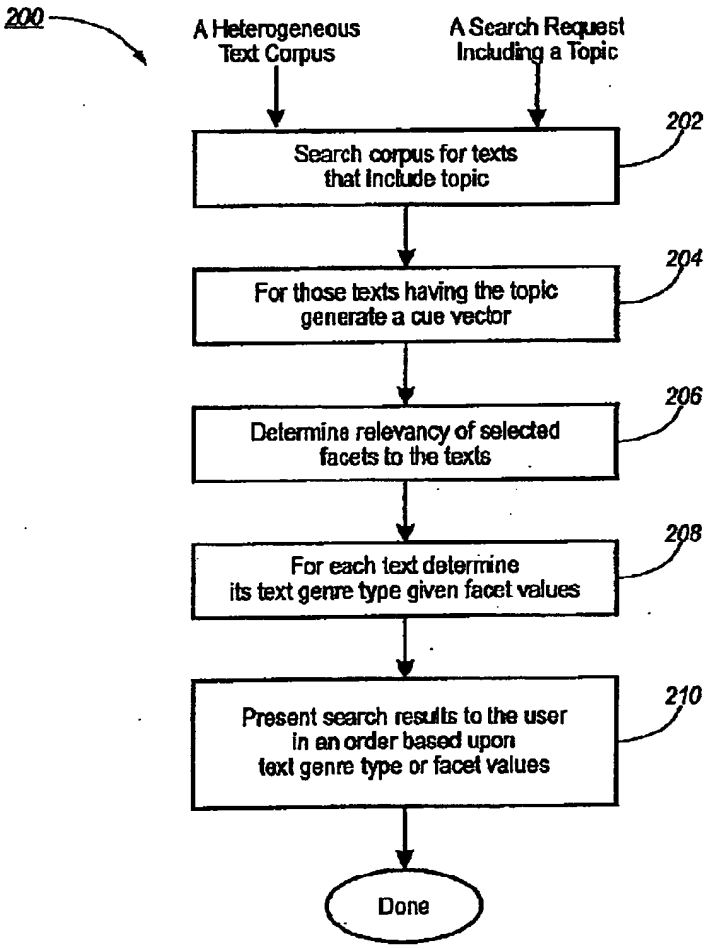
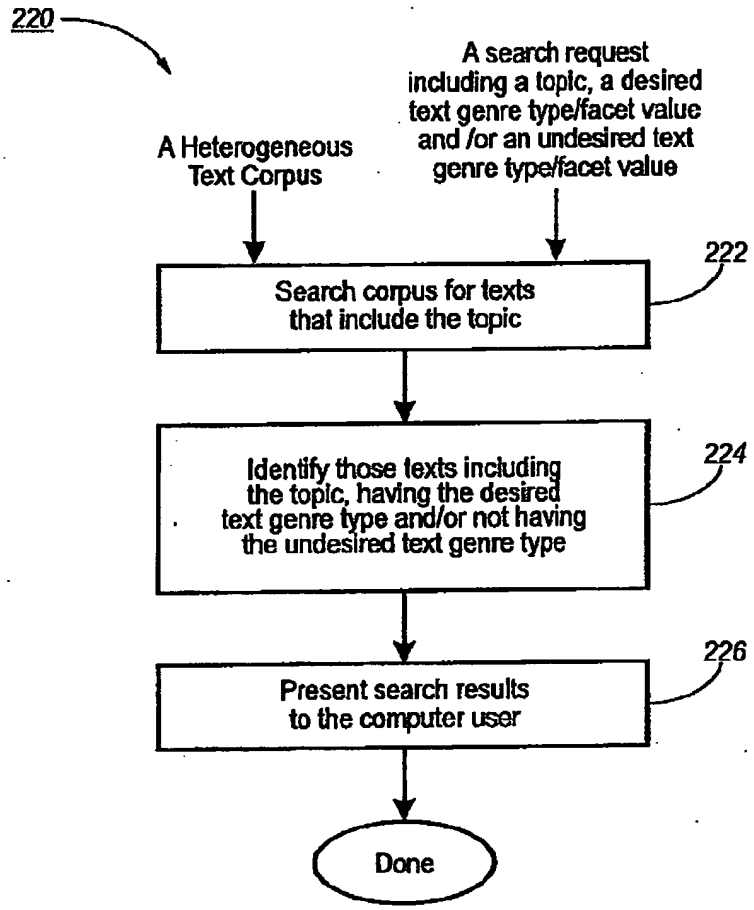

200
A Heterogeneous Text Corpus
A Search Request Including a Topic
Search corpus for texts that include topic
202
For those texts having the topic generate a cue vector
204
Determine relevancy of selected facets to the texts
206
For each text determine its text genre type given facet values
208
Present search results to the user in an order based upon text genre type or facet values
210
Done

FIG.6


220
A search request including a topic, a desired text genre type/facet value and /or an undesired text genre type/facet value
A Heterogeneous Text Corpus
222
Search corpus for texts that include the topic
224
Identify those texts including the topic, having the desired text genre type and/or not having the undesired text genre type
226
Present search results to the computer user
Done

1 Abstract

A processor impremented method of identifying the genre of a machine readable, untagged text. The processor implemented method begins by generating a cue vector form the text which represents occurrences in the text of a first set of nonstructural, surface cues, which are easily computable. Afterward, the processor dctermines whether the text is an instance of a first text genre using the cue vector and a weighting vector associated with the first text genre.

2 Representative Drawing

Figure 3